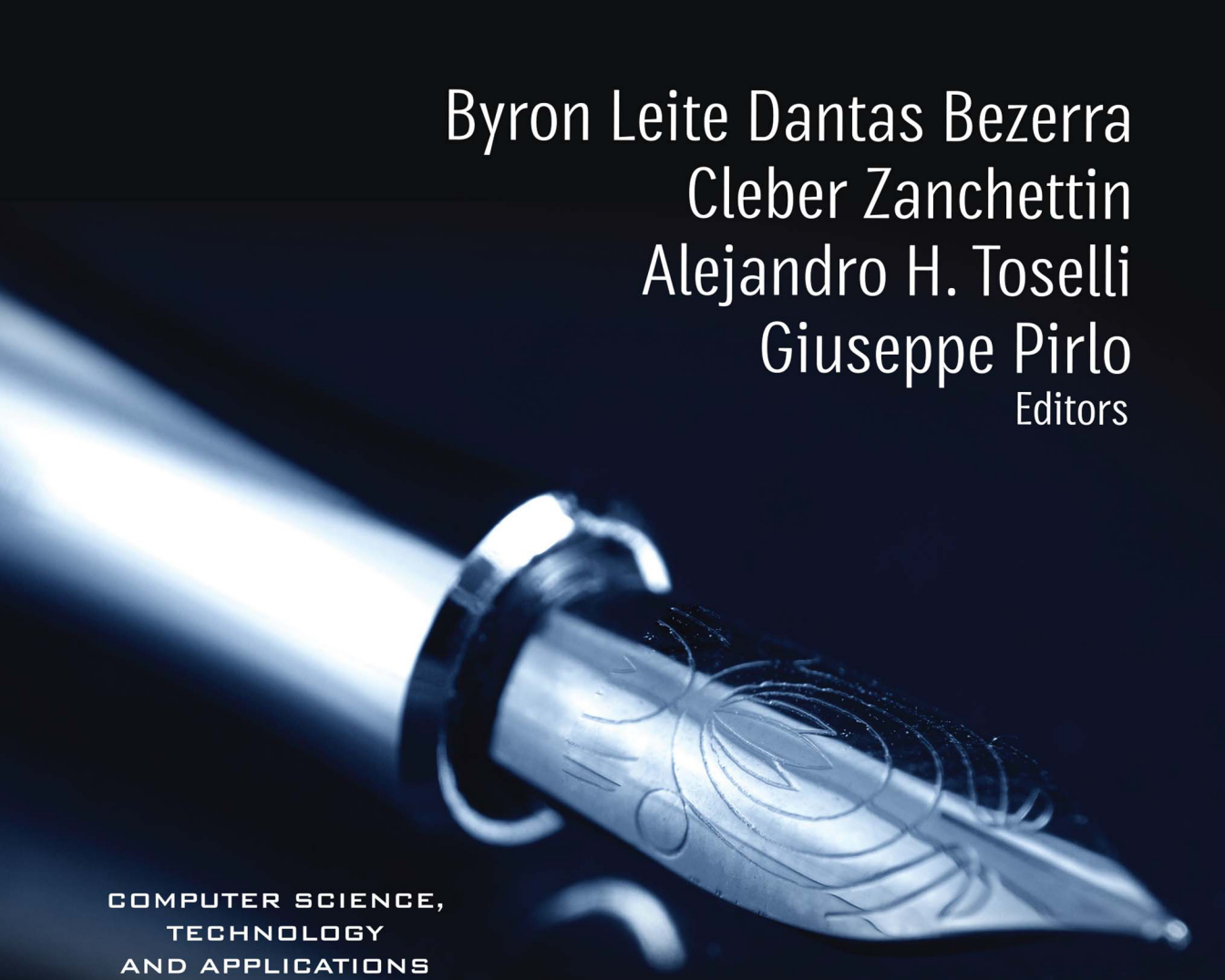
Byron Leite Dantas Bezerra
Cleber Zanchettin
Alejandro H. Toselli
Giuseppe Pirlo
Editors



COMPUTER SCIENCE,
TECHNOLOGY
AND APPLICATIONS

Handwriting

Recognition, Development and Analysis

NOVA

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Page 1 of 404
Complimentary Contributor Copy

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 2 of 404

HANDWRITING

RECOGNITION, DEVELOPMENT AND ANALYSIS

No part of this digital document may be reproduced, stored in a retrieval system or transmitted in any form or by any means. The publisher has taken reasonable care in the preparation of this digital document, but makes no expressed or implied warranty of any kind and assumes no responsibility for any errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of information contained herein. This digital document is sold with the clear understanding that the publisher is not engaged in rendering legal, medical or any other professional services.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

COMPUTER SCIENCE, TECHNOLOGY AND APPLICATIONS

Additional books in this series can be found on Nova's website under the Series tab.

Additional e-books in this series can be found on Nova's website under the eBooks tab.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

COMPUTER SCIENCE, TECHNOLOGY AND APPLICATIONS

HANDWRITING

RECOGNITION, DEVELOPMENT AND ANALYSIS

BYRON LEITE DANTAS BEZERRA
CLEBER ZANCHETTIN
ALEJANDRO H. TOSELLI
AND
GIUSEPPE PIRLO
EDITORS



Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

Copyright © 2017 by Nova Science Publishers, Inc.

All rights reserved. No part of this book may be reproduced, stored in a retrieval system or transmitted in any form or by any means: electronic, electrostatic, magnetic, tape, mechanical photocopying, recording or otherwise without the written permission of the Publisher.

We have partnered with Copyright Clearance Center to make it easy for you to obtain permissions to reuse content from this publication. Simply navigate to this publication's page on Nova's website and locate the "Get Permission" button below the title description. This button is linked directly to the title's permission page on copyright.com. Alternatively, you can visit copyright.com and search by title, ISBN, or ISSN.

For further questions about using the service on copyright.com, please contact:
Copyright Clearance Center
Phone: +1-(978) 750-8400 Fax: +1-(978) 750-4470 E-mail: info@copyright.com.

NOTICE TO THE READER

The Publisher has taken reasonable care in the preparation of this book, but makes no expressed or implied warranty of any kind and assumes no responsibility for any errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of information contained in this book. The Publisher shall not be liable for any special, consequential, or exemplary damages resulting, in whole or in part, from the readers' use of, or reliance upon, this material. Any parts of this book based on government reports are so indicated and copyright is claimed for those parts to the extent applicable to compilations of such works.

Independent verification should be sought for any data, advice or recommendations contained in this book. In addition, no responsibility is assumed by the publisher for any injury and/or damage to persons or property arising from any methods, products, instructions, ideas or otherwise contained in this publication.

This publication is designed to provide accurate and authoritative information with regard to the subject matter covered herein. It is sold with the clear understanding that the Publisher is not engaged in rendering legal or any other professional services. If legal or any other expert assistance is required, the services of a competent person should be sought. FROM A DECLARATION OF PARTICIPANTS JOINTLY ADOPTED BY A COMMITTEE OF THE AMERICAN BAR ASSOCIATION AND A COMMITTEE OF PUBLISHERS.

Additional color graphics may be available in the e-book version of this book.

Library of Congress Cataloging-in-Publication Data

Names: Bezerra, Byron Leite Dantas, editor.
Title: Handwriting : recognition, development and analysis / editors, Byron Leite Dantas Bezerra, Cleber Zanchettin, Alejandro H. Toselli and Giuseppe Pirlo (Department of Computer Engineering, University of Pernambuco, Recife, Brazil, and others).
Description: Hauppauge, New York : Nova Science Publishers, Inc., [2017] | Series: Computer science, technology and applications | Includes bibliographical references and index.
Identifiers: LCCN 2017019936 (print) | LCCN 2017021182 (ebook) | ISBN 9781536119572 (eBook) | ISBN 9781536119374 (hardcover) | ISBN 9781536119572 (Ebook)
Subjects: LCSH: Optical character recognition devices. | Graphology--Data processing. | Writing--Identification--Data processing. | Pen-based computers.
Classification: LCC TA1640 (ebook) | LCC TA1640 .H36 2017 (print) | DDC 006.4/25--dc23
LC record available at https://lccn.loc.gov/2017019936

Published by Nova Science Publishers, Inc. † New York

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

CONTENTS

Preface		**vii**
Part I	**Recognition and Development**	**1**
Chapter 1	Handwriting Recognition: Overview, Challenges and Future Trends *Everton Barbosa Lacerda, Thiago Vinicius M. de Souza, Cleber Zanchettin, Juliano Cícero Bitu Rabelo and Lara Dantas Coutinho*	**3**
Chapter 2	Thresholding *Edward Roe and Carlos Alexandre Barros de Mello*	**33**
Chapter 3	Historical Document Processing *Basilis Gatos, Georgios Louloudis, Nikolaos Stamatopoulos and Giorgos Sfikas*	**57**
Chapter 4	Wavelet Descriptors for Handwritten Text Recognition in Historical Documents *Leticia M. Seijas and Byron L. D. Bezerra*	**95**
Chapter 5	How to Design Deep Neural Networks for Handwriting Recognition *Théodore Bluche, Christopher Kermorvant and Hermann Ney*	**113**
Chapter 6	Handwritten and Printed Image Datasets: A Review and Proposals for Automatic Building *Gearlles V. Ferreira, Felipe M. Gouveia, Byron L. D. Bezerra, Eduardo Muller, Cleber Zanchettin and Alejandro Toselli*	**149**
Part II	**Analysis and Applications**	**167**
Chapter 7	Mathematical Expression Recognition *Francisco Álvaro, Joan Andreu Sánchez and José Miguel Benedí*	**169**
Chapter 8	Online Handwriting Recognition of Indian Scripts *Umapada Pal and Nilanjana Bhattacharya*	**211**

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

Chapter 9 Historical Handwritten Document Analysis of Southeast Asian Palm Leaf Manuscripts **227**
Made Windu Antara Kesiman, Jean-Christophe Burie, Jean-Marc Ogier, Gusti Ngurah Made Agus Wibawantara and I Made Gede Sunarya

Chapter 10 Using Speech and Handwriting in an Interactive Approach for Transcribing Historical Documents **277**
Emilio Granell, Verónica Romero and Carlos-D. Martínez-Hinarejos

Chapter 11 Handwritten Keyword Spotting the Query by Example (QbE) Case **297**
Georgios Barlas, Konstantinos Zagoris and Ioannis Pratikakis

Chapter 12 Handwriting-Enabled E-Paper Based on Twisting-Ball Display **321**
Yusuke Komazaki and Toru Torii

Chapter 13 Speed and Legibility: Brazilian Students Performance in a Thematic Writing Task **333**
Monique Herrera Cardoso and Simone Aparecida Capellin

Chapter 14 Datasets for Handwritten Signature Verification: A Survey and a New Dataset, the RPPDI-SigData **345**
Victor Kléber Santos Leite Melo, Byron Leite Dantas Bezerra, Rebecca H. S. N. Do Nascimento, Gabriel Calazans Duarte de Moura, Giovanni L. L. de S. Martins, Giuseppe Pirlo and Donato Impedovo

Chapter 15 Processing of Handwritten Online Signatures: An Overview and Future Trends **363**
Alessandro Balestrucci, Donato Impedovo and Giuseppe Pirlo

Editor's Contact Information **387**

Index **391**

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

PREFACE

This book has the primary goal to present and discuss some recent advances and ongoing developments in the Handwritten Text Recognition (HTR) field, resulting from works done on different HTR-related topics for the achievement of more accurate and efficient recognition systems. Nowadays, there is an enormous worldwide interest in HTR systems, which is mostly driven by the emergence of new portable devices incorporating handwriting recognition functions. Others interests are the biometric identification systems employing handwritten signature, as well as the requirements from cultural heritage institutions like historical archives and libraries in order to preserve their large collections of historical (handwritten) documents. The book is organized into two sections: the first one is mainly devoted to describing the current state-of-the-art in HTR and the last advances in some of the steps involved in HTR workflow (that is, preprocessing, feature extraction, recognition engines, etc.), whereas the second focuses more on some relevant HTR-related applications.

In more depth, the first part offers an overview of the current state-of-the-art of HTR technology and introduces the new challenges and research opportunities in the field. Besides, it provides a general discussion of currently ongoing approaches towards solving the underlying search problems on the basis of existing methods for HTR in terms of both accuracy and efficiency. In particular, there are chapters especially focused on image thresholding and enhancement, text image preprocessing techniques for historical handwritten documents and feature extraction method for HTR. Likewise, in line with the breakout success of Deep Neural Networks (DNNs) in the field, a whole chapter is devoted to describing the designing of HTR systems based on DNNs. Finally, a chapter listing the most used benchmarking datasets for HTR is also included, providing detailed information about which types of HTR systems (on/off-line) and features are commonly considered for each of them.

In the second part, several systems – also developed on the basis of the fundamental concepts and general approaches outlined in the first part – are described for several HTR-related applications. Presented in the corresponding chapters, these applications cover a wide spectrum of scenarios: mathematical formulae recognition, scripting language recognition, multimodal handwriting-speech recognition, hardware design for on-line HTR, student performance evaluation through handwriting analysis, performance evaluation methods, keyword spotting, and handwritten signature verification systems.

Last but not least, it is important to remark that to a large extent, this book is the result of works carried out by several researchers in the Handwritten Text Recognition field.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

Therefore it owes credit to these researchers that have directly contributed to their ideas, discussions and technical collaborations and in general who, in one manner or another, have made it possible.

January 31st, 2017
The Editors

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 10 of 404

PART I.
RECOGNITION AND DEVELOPMENT

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

In: Handwriting: Recognition, Development and Analysis ISBN: 978-1-53611-937-4
Editors: Byron L. D. Bezerra et al. © 2017 Nova Science Publishers, Inc.

Chapter 1

HANDWRITING RECOGNITION: OVERVIEW, CHALLENGES AND FUTURE TRENDS

***Everton Barbosa Lacerda*[1,*], *Thiago Vinicius M. de Souza*[1,†], *Cleber Zanchettin*[1,‡], *Juliano Cícero Bitu Rabelo*[2,§] *and Lara Dantas Coutinho*[2,¶]**
[1]Centro de Informática,
Universidade Federal de Pernambuco, Recife, Brazil
[2]Document Solutions, Recife, Brazil

1. Introduction

Handwriting recognition emerged as an important research field since the early days of computer science and engineering development. Furthermore, the appealing motivation and convenience of automatically reading our paper documents and converting them to digital format have always pushed the area forward. Both academia and industry have been developing studies and products which aim to read digital documents. Besides, in spite of major efforts devoted to bring out a paper-free society, a huge number of paper documents are generated and processed by computers every day, all over the world, in order to handle, retrieve and store information (Bortolozzi et al., 2005).

At the beginning, due to several aspects, machine printed documents have evolved quicker and sooner. This fact is the result of the constrained set of symbols (the available fonts in computer systems) and their uniform layout, size, and position. In addition to this, structured layouts are commonly seen in machine documents, which also facilitates the process of recognition, since it makes the finding and isolation of words or characters easier. Therefore, the increasing use and dissemination of OCR software are based on structured printed documents.

*E-mail address: ebl3@cin.ufpe.br.
†E-mail address: tvms@cin.ufpe.br.
‡E-mail address: cz@cin.ufpe.br.
§E-mail address: jcbr@documentsolutions.com.br.
¶E-mail address: ldc@documentsolutions.com.br.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

The research related to handwritten characters recognition began in the 1950s with the creation of the first commercial OCRs. Even with the technological advances of imaging processing and acquisition devices, the scenario still remains challenging and current for new researchers. The task itself consists of detecting and recognizing the characters present in an input image and converting them to the binary pattern, corresponding to the character found.

The character recognition process is handled according to how the writing information is obtained. There are two most common ways of obtaining information about the writing of characters: (1) when there are pre-existing handwritten documents and the input images are acquired via scanners or photo cameras, the process is called ”offline recognition”. In this scenario, we only have information about the image intensity, that is, the values of each pixel in the coded image; (2) when the writing is directly made in devices capable of capturing the Cartesian coordinates and information inherent to the writing process itself, such as stroke velocity, pen pressure or the order of the traces, the process is called ”online recognition”. Generally, the effective use of that temporal information by online recognition techniques yields better results in comparison to offline methods.

The recognition of manuscripts is much harder than that of printed texts. Several factors contribute to this: (i) the great variability of writing styles, leading to a virtually infinite set of possible formats for the same symbol or letter; this is easily seen on the writing of different people, but also happens when a person’s calligraphy change over time; (ii) the similarity between some characters is high; (iii) touching and overlapping characters (Mello et al., 2012). Poorly written and degraded characters may turn the recognition of this kind of document even more difficult.

The aforementioned issues refer only to characters themselves. However, besides those, there are difficulties that affect both printed and handwritten documents, such as background noise, poor image quality, degradation over time, etc. However, normally, those problems appear to be more hazardous for handwriting recognition. Maybe that results from the intrinsic hardness to cope with this kind of document as opposed to printed text; it is not possible to make general assumptions about the document content or layout, which may facilitate the recognition process.

Regarding the text recognition, there are some strategies that refer to text granularity, i.e., whether we are concerned about sentences, words, or characters. The classical approach is to segment the document into regions, lines, words, and finally, characters, and execute the classification of symbols, which correspond to some alphabet, e.g. Latin, Arabic, Chinese, etc. In that case, the classification phase aims to label each isolated character to its correct class.

By far, most of the applications and research in recognition are based on that framework. Nevertheless, because of some hindrances of isolating characters, there are also methods working on words or sentences. The advantages are the possibilities of using context, in other words, utilize the results of previous words to assist the recognition of the next one; or the application of dictionaries which can help to correct the words, in the case of some wrong characters.

Thereat, in this book we address different fields and challenges of handwriting recognition. In doing this, we choose to divide it into two main parts: the first part of the book, called Recognition and Development, comprising Chapters 1–6, covers core concepts and

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 14 of 404

challenges.

In this first chapter, we begin presenting the most recent methods in each of the mentioned approaches. In order to make understanding and reading easier, and besides, to enable specific search, i.e., to make possible the consultation of the desired domain only, we illustrate the different application areas (digits, characters and words) separately, in the following sections. Later, we comment about the tendencies and possible future outbreaks in this evolving and fascinating research field.

Chapter 2 explores some recent algorithms for thresholding document images. Although this is a theme with works dated from decades ago, it is still unsolved. When documents have particular features as texture, patterns, smears, smudges, folding marks, adhesive tape marks, ink-bleed or bleed through effect, the process of finding the correct separation between background and foreground is not so simple.

In Chapter 3, the recent advances and ongoing developments for historical handwritten document processing are investigated. It outlines the main challenges involved, the different tasks that have to be implemented, as well as practices and technologies that currently exist in the literature.

Chapter 4 investigates different approaches for feature extraction, revising the literature and proposing an approach based on the application of the CDF 9/7 Wavelet Transform (WT) in order to represent the content of each slice.

Chapter 5 revises important aspects to take into account when building neural networks for handwriting recognition in the hybrid NN/HMM framework, providing a better understanding and evaluation of their relative importance. The authors show that deep neural networks produce consistent and significant improvements over networks with one or two hidden layers, independently of the kind of neural network (MLP or RNN) and of input (handcrafted features or pixels).

Motivated by: (i) the absence of datasets available for every language used in the world; (ii) none of the existent datasets for a specific language is large and diverse enough to produce recognition systems as reliable as human readers; (iii) manually building large image text datasets can be impractical if we take into account the diversity of applications in the real world; Chapter 6 presents two techniques to generate large and diverse datasets, one for handwritten image texts and other for machine printed ones.

In the second part of this book, named Analysis and Applications, Chapters 7–15, different authors propose techniques to address with handwriting recognition in diverse contexts. In Chapter 7, the authors present the main challenges in the recognition of mathematical expressions, and propose an integrated approach to address them. A formal statistical framework of a model based on two-dimensional grammars and its associated parsing algorithm are presented.

Chapter 8 presents the state of the art of online handwriting recognition of main Indian scripts and then proposes a general scheme to recognize Indian scripts. The authors combine online and offline information to classify segmented primitives.

Chapter 9 describes in detail the historical handwritten document analysis for Southeast Asian palm leaf manuscripts by reporting the latest studies and experimental results of document analysis tasks which range from corpus collection, ground truth data generation, binarization process to the isolated character recognition and the word spotting tasks.

A multimodal interactive transcription system where user feedback is provided by

Cerence Operating Company, Exhibit 1017
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 15 of 404

means of touchscreen pen strokes, traditional keyboard, and mouse operations is presented in Chapter 10. The combination of the main and the feedback data streams is based on the use of Confusion Networks derived from the output of three recognition systems: two handwritten text recognition systems (*off-line* and *on-line*), and an automatic speech recognition system.

Chapter 11 exploits the evolution of keyword spotting in handwritten documents focusing on the Query by Example case where the query is a word image. It aims to present in a concise manner the distinct algorithms which have been presented for over two decades so that useful conclusions should be drawn for the future steps in this exciting research area.

The details of the development, including background, structure, fabrication method, performance and applications of handwriting-enabled twisting-ball display is discussed in Chapter 12. This technology will be applicable to next-generation of electronic whiteboard.

Proficiency of writing skills is even a goal that students should achieve. In this context, Chapter 13 aims to investigate the performance of Brazilian students during the thematic writing task regarding the speed and legibility criteria, according to the Brazilian adaptation of the Detailed Assessment of Speed of Handwriting. Although this study is not directly related to automatic reading methods or practices, it traverses about the object of recognition methods: handwriting; and therefore, it is interesting to consider the writing process when thinking about automatic reading.

Other interesting application of handwriting recognition is the automatic processing of signatures. In this scenario, the purpose of Chapter 14 is to analyze and discuss the most used data sets in the literature in order to find what are the challenges pursued by the community in the past few years. In addition to this, they propose a new dataset and appropriated metrics to analyze and evaluate signature verification techniques.

The possibility of acquiring handwritten on-line signatures is exponentially rising due to the availability of low-cost acquisition systems integrated into mobile devices, such as smartphones, tablets, PDAs, etc. In Chapter 15, the most interesting current challenges of the handwritten on-line signature processing are identified and promising directions for further research are highlighted.

2. Models

This section presents a brief overview and explanation of the models that lay the foundation for state-of-the-art methods. It is not meant to explore all details about all algorithms development and training, however, it is possible to understand their principles and ideas, which in this context is sufficient to ease the understanding of literature techniques, and may help to select one or other algorithm when developing new methods.

2.1. K-Nearest Neighbors

K-Nearest Neighbors (k-NN) is one of the most used and simplest algorithms in machine learning. The underlying idea is: similar examples tends to be ”near” to each other when one thinks about their characteristics. Or in other words, if a sample belongs to a certain class, it is expected that features values of another example which also belongs to the same

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 16 of 404

class do not deviate so much from the values of the former. Thus, the distance between those instances should be small.

Therefore, k-NN is a nonparametric method which works as follows: we store all available samples, that correspond to the training data; it is composed of those examples features and their labels. Then, we compare the input example to all training data and assign its class to be the same of the majority of k nearest samples of the training set. k is a user-defined constant. Figure 1 shows the functioning principle of k-NN. In this scenario, the input is marked as a star, while we have two classes (empty and full circles). In both situations illustrated by Figure 1, $k = 3$ and $k = 7$, the method predicts the input as belonging to the full circle class.

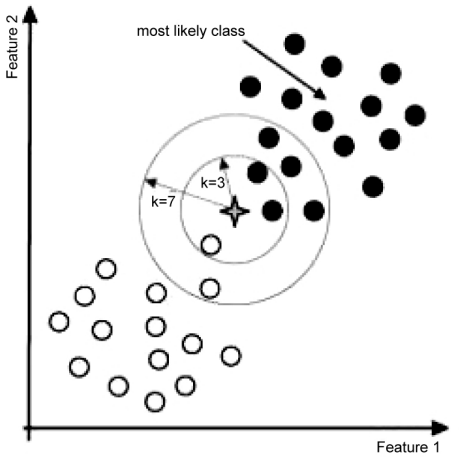


Figure 1. k-NN operating idea.

Thus, it is possible to observe that parameter k plays a pivotal role in k-NN results. It is not hard to notice that depending on this value, the algorithm may change its prediction. So, the best value for this parameter depends on the problem, and specifically, on the data. A common practice for the estimation of k value is to vary the value from one to the square root of the number of training samples (Duda et al., 2000), and choose the value that achieved best results. Alternatively, it is possible to weight the neighbors such that nearer neighbors contribute more to the result. In that case, the weights are normally proportional to the inverse of the distance from the query point.

Proximity definition depends on the distance measure. The most employed distance metric in k-NN is Euclidean distance, although it is possible to find various others in the literature, such as city-block or Manhattan, Mahalanobis, Minkowski, only to cite a few. As k, the distance measure may also change output values of the algorithm. An exploratory analysis can indicate the most suitable measure for a specific data set.

The shortcomings of k-NN are mainly the dependency of data structure, and the great processing cost overload when the training set increases. Since the algorithm is based on a comparison of the input to the training data, the greater the training set, the greater is the number of operations, and, consequently, also is the processing time. To overcome this, it is possible to implement pruning policies, which aims to decrease the number of training examples, normally based on some similarity measure. In this context, it is known that

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 17 of 404

several similar examples probably do not help to discriminate data. Therefore, some of those examples could be excluded without generalization loss, thus helping to improve the performance of the method.

2.2. Multilayer Perceptrons

Multilayer Perceptrons (MLP) are feed-forward networks which have one or more hidden layers, usually composed of sigmoidal activation function neurons. Interesting properties of MLP are that with one hidden layer, it can approximate any continuous function (Cybenko, 1989), while using two hidden layers allows the approximation of any function (Duda et al., 2000). Figure 2 shows a schematic view of an MLP net with three nodes in the input layer, one hidden layer with four nodes, and an output layer of two nodes.

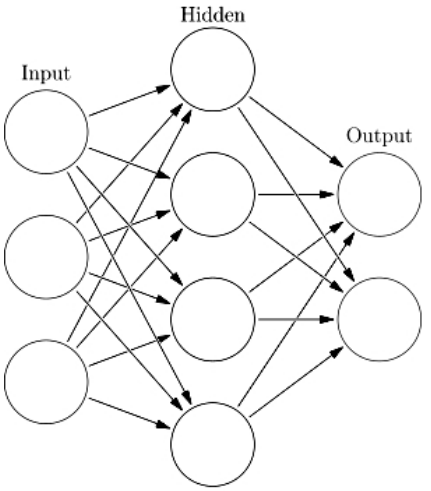


Figure 2. Schematic visualization of MLP.

Initially, neural networks were formed by one layer, and consequently, their training was straightforward, since the output is directly observable, and thus, can be used to guide weights adjusting. The more drastic fact is: single layer networks were only able to solve linearly separable problems. The solution to that limitation appeared with the description of the backpropagation algorithm (Rumelhart et al., 1986a)(Rumelhart et al., 1986b).

The fundamental idea of the algorithm is to use gradient descent to calculate hidden layers errors by an estimate of the effect caused by them over output layer error. Thus, the output error is calculated and is backpropagated to hidden layers, making possible weights updating proportionally to the values of connections between the layers. Due to the use of gradient, activation functions need to be differentiable. That justifies the use of sigmoid functions, since they are a differentiable approximation to step function (early used in Rosenblatt's Perceptron (Haykin, 2009)).

The training proceeds in two phases: (i) *forward phase*, when the input signal is propagated through the network until it reaches the output; (ii) *backward phase*, when an error is obtained by comparing the output of the network with the desired response. That resulting error is propagated from output to input (backward direction), layer by layer, allowing the

Cerence Operating Company, EX-2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 18 of 404

adjustments of weights of the network.

MLP is one of the most used and studied machine learning techniques over the world (in addition to its classical training strategy, the backpropagation algorithm). That also happens for recognition in general, including the special case of handwriting. Thus, there are several studies where one can find all minutiae about MLP training, including the mathematical derivation which gives exact formulae to weights updating (by chain rule) (Haykin, 2009)(Braga et al., 2007). Furthermore, it is also possible to read about MLP's various other training algorithms, such as Quickprop (Fahlman, 1988), R-prop (Riedmiller and Braun, 1993), Levenberg-Marquardt (Hagan and Menhaj, 1994), to name a few. Those algorithms arose to overcome known difficulties encountered by classical backpropagation algorithm, such as slow convergence, sensibility to local minima, etc. More details about that method are presented in Chapter 5.

2.3. Support Vector Machines

Support Vector Machine (SVM) (Cortes and Vapnik, 1995) is a binary machine learning method based on statistical learning theory (Vapnik, 2000), with some highly elegant properties. The main idea may be summed up as follows: given a training sample, the support vector machine constructs a hyperplane as the decision surface in such a way that the margin of separation between positive and negative examples is maximized (Haykin, 2009). This feature contrasts with other learning techniques which encounter any separation surface, such as MLP or RBF, for instance. The margin of separation is defined as the smaller distance between training patterns and the decision surface, which in this situation, may be referred as the optimal hyperplane. Figure 3 illustrates the difference between an arbitrary decision surface with a smaller margin (item 2.3), and an optimal hyperplane, which possesses the maximal margin of separation (item 2.3).

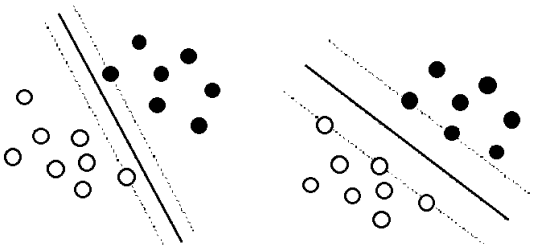
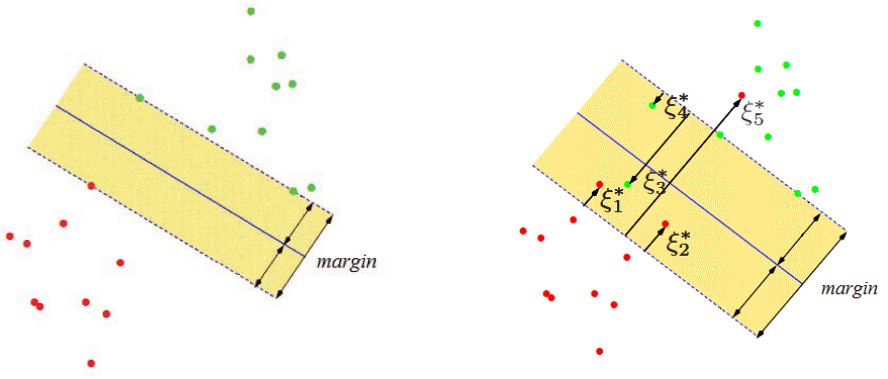


Figure 3. Correct decision surfaces: (a) smaller margin, and (b) maximal margin.

The basic procedure to determine the optimal hyperplane only permits the classification of linearly separable data. In order to treat non-separable data, it is introduced the concept of soft margins. In that case, classification errors are allowed in training, to provide wider margins, which tends to augment generalization power of the classifier. Figure 4 shows this situation, exhibiting both linearly separable and non-separable data in items 2.3 and 2.3 respectively, where the points marked as ξ are in the wrong side of the decision surface.

Although margin softening is quite useful, it is not self-sufficient to give the required

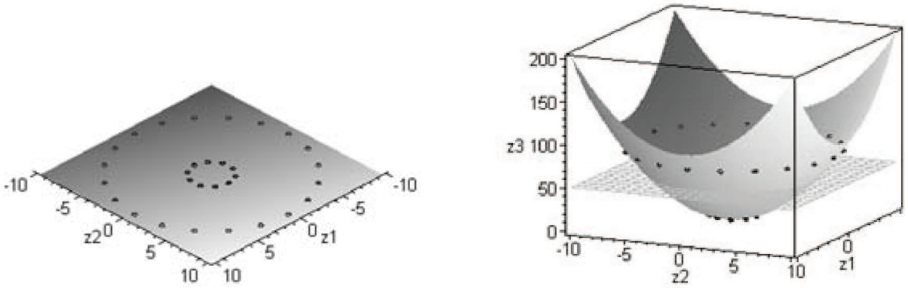
Cerence Exhibit 1010
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 19 of 404



(a) (b)

Figure 4. Support vectors classification: (a) linearly separable data; hard margins, and (b) non linearly separable data; soft margins (adapted from (Hastie et al., 2009)).

classification skills to SVM. This accrues from the fact the hyperplanes are not always adequate to separate input data; there are situations in which a non-linear surface would be more suitable. This conversion or mapping is obtained by the ”kernel trick” (Haykin, 2009). The idea is: non-linearly separable input data may become linearly separable in other hyperspace, in which it will be possible to define a hyperplane that discriminates the given data. Figure 5 presents such a transformation.



(a) (b)

Figure 5. Kernel mapping: (a) input space, (b) kernel space.

More details about SVM training, including the optimization problem to determine support vectors can be consulted in (Haykin, 2009)(Hastie et al., 2009). And, as others classifiers, SVM also has parameters which define its performance at a certain problem. The main ones are the kernel function and its internal parameters: as example of functions, we can cite polynomial and Gaussian or RBF (radial basis function); and the regularization constant, used on the margins softening (Hastie et al., 2009). By the way, SVM are more sensible to the variation of its parameters, which is considered a shortcoming of the method (Braga et al., 2007).

As exposed, SVM is originally built to treat binary problems. Naturally, there are several multiclass problems; at this point, two principal strategies have been used to deal with

Exhibit 1002
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 20 of 404

this situation. One of them is based on modifying SVM training; interesting results were obtained, but, the computational cost is very high. The second approach consists of the decomposition multiclass scenario in various binary groups, where SVM are applied as usually. This strategy is used more often (Braga et al., 2007).

2.4. Committees

Generally, the most common practice in the use of learning machines is to perform several trainings with a set of examples, testing the performance of the model in a validation set, proceeding by modifying the model parameters, until obtaining a better performance and finally using it in the test set to get the best hit rate. This approach makes us think that we are choosing the best possible classifier. However, it is worth mentioning that there is a large stochastic factor when selecting validation sets, and even with a carefull distribution of this set, it is possible that the network does very well on that chosen part but does not have the same performance with the test set.

Machine learning committees are mechanisms that seek to combine the performance of several specialist machines to achieve the same common goal. The idea is that the weakness of a machine trained in a particular situation is compensated by the support of the others (Sammut and Webb, 2011). Those mechanisms are generally constructed in two ways, being defined as static or dynamic structures.

Static structure committees use a mechanism that combines the result of several classifiers without the interference of input signals. The ensemble is a committee of static structure that performs a linear combination of different machine results, which can be done by performing the average of the outputs of the machines or selecting the best result among the best votes. Another example of a static committee is the boosting mechanism that combines several weak classifiers into a strong classifier. The Adaboost (Freund et al., 1996) algorithm, is a remarkable example that represents this type of mechanism. In this technique, the idea is to train a series of new and stronger classifiers to correct the mistakes made by previous machines and to combine the output results.

In a dynamic structure committee, the input signal acts directly on the mechanism that combines the output results. One of the most common models of dynamic structure committees is called a mixture of experts (Jacobs et al., 1991). In that model, several classifiers are trained in different regions of the input data. The switching between the regions of the input data and the models to be used in those regions is carried out through the interaction of the input signal.

2.5. Deep Learning

Deep learning is an area of artificial intelligence that studies algorithms that learn from experience and understand the world from a hierarchy of concepts, where each concept is defined in terms of its relations with simpler concepts. By referencing knowledge from experience, this approach avoids the need for human interaction to formally specify all the knowledge the machine needs. The concept hierarchy allows the computer to learn complicated concepts from constructions that use simpler concepts. If a graph is assembled to show how those concepts are built on top of other concepts, this graph would be deep,

Cerence Operating Company Exhibit 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 21 of 404

with many layers. For this reason, these approaches are called Deep Learning (Goodfellow et al., 2016).

Today the field of research in Deep Learning is extremely heated mainly because techniques of this area are achieving the best results regarding the tasks of classification, detection, and localization of images, processing of natural languages, speech and audio. Within the models known in the Deep Learning scenario, the convolutional and Long-Shot-Term-Memory (LSTM) networks have been obtaining the best results in most fields of research in machine learning. Today big companies like Google, Baidu, Facebook and others use these types of models in their main systems.

The concept of Deep Learning involves a cascade of non-linear transformations, using end to end learning, with supervised, unsupervised, probabilistic approaches and normally hierarchical. The following sections will briefly describe the operation of one of the most common models in handwriting recognition. A more detailed revision can be found in Goodfellow et al. (2016) and is also presented in Chapter 5.

2.6. Convolutional Neural Networks

Convolutional Neural Networks or CNNs (LeCun and Bengio, 1995) are a type of neural network where at least one of its layers is composed of a characteristic extractor based on convolution operations. This type of network is generally used for sorting, detecting, and locating objects of interest where input data is in mesh format or more specifically structured in arrays. One of the great advantages of CNNs in comparison to traditional strategies is the sharing of weights for the different regions of the input data, enabling an improvement in learning through the detection of local characteristics present in different parts of the matrix. Convolutional networks today represent the state of the art in most image recognition tasks. The popularization of this type of network started with the good results shown in (Krizhevsky et al., 2012).

CNNs are inspired by biology and neuroscience, as they rely heavily on the functioning of the visual cortex. Hubel and Wise (1962) conducted experiments where they proved that specific cells of the visual cortex are activated when edges were displayed in a certain orientation. In other words, different parts of the cortex specialize in recognizing a particular type of feature and work together to recognize the object as a whole (Hubel and Wiesel, 1962), so CNNs are designed to work similarly.

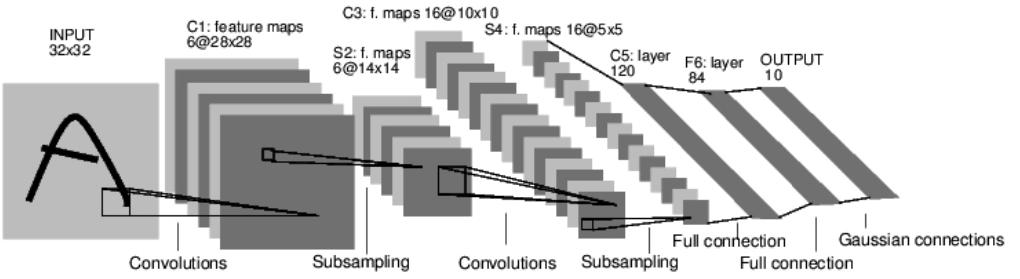


Figure 6. Convolutional Network (LeCun and Bengio, 1995).

Like MLP networks, convolutional networks are formed by several layers. Those layers

Cerence Operating Company, Ex. 2014
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 22 of 404

may be arranged to sequentially perform similar or different functions. The first layer of a CNN is usually the convolution layer. That layer is responsible for receiving the input image with dimensions N1xM1xD. In it, the convolution operations are carried out with the aid of a filter with dimensions N2xM2xD, where the weights of the network are present. This operation results in a map representing the extracted features of the image.

In CNNs, taking into account the space domain, the convolution operation consists in carrying out the displacement of a mask of weights throughout the image obeying a certain orientation and direction where, for each position of the displacement, the internal product between the elements of the mask and the elements of the image region below the mask is calculated. At each offset, an activation value is generated that will be assigned to the feature map at the position relative to the center of the mask over the image. A depth level of the feature filter used represents a depth level in the resulting feature map. In that first step, the resulting feature map tends to specialize in low-level features found in image objects, such as small edges and curves. In the next stage of execution of the network, the map of features is passed as input to other layers, where representations will gain new levels of abstractions.

In addition to the convolution layer, CNNs have other types of layers as the network becomes deeper. Generally, a convolution layer is accompanied by an activation layer that limits the values passed as input to the other layers. The most commonly used activation function in convolutional networks and the Rectified Linear Unit (ReLu) (Glorot et al., 2011). The function brought interesting results in the training phase and in the prevention of overfitting, compared to previously used functions such as the hyperbolic tangent and the sigmoidal function.

Another common and important layer within the CNN universe is the Max-Pooling (Huang et al., 2007) layer that promotes a downsampling of the input data by selecting only the largest value within a neighborhood of the image. Pooling helps render representations of images more invariant to operations such as translation in input data.

The last layer of a network is usually a fully-connected layer where the feature maps are concatenated into a single vector and passed to the layer that will return the probabilities of the instance belonging to a previously trained class. Training of a convolutional network is performed using the backpropagation algorithm (Rumelhart et al., 1988). In that algorithm, the result of the classification of the network is compared with the label of the example class, then the classification error is retropropagated towards the previous layers so that their weights are updated.

2.7. Long-Short Term Memory

Recurrent Neural Networks are a family of networks used to treat sequential data. For this type of data, the network parameters are shared between different time periods. Recurring neural networks, in theory, should handle well problems with sequences of any size, from the very long to the shortest. However, in practice, this is not what happens; this problem must be due to the vanishing gradient problem (Bengio et al., 1994).

The Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) is a Recurrent Neural Network model that introduces a new structure called "memory cell" to address the problem of the vanishing gradient. Because it is a recurrent network, its archi-

Cerence Operating Company, IPR2024-01267
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 23 of 404

tecture is very similar to the other models of this family, differentiating itself precisely by the use of the memory cell structure.

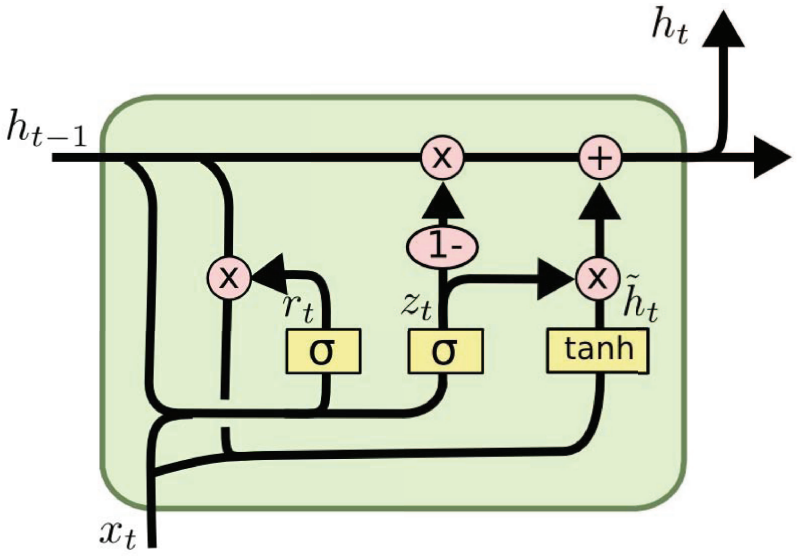


Figure 7. Visualization of a Memory Cell (LeCun and Bengio, 1995).

A memory cell is composed of elements that allow access and regulate the sharing of information between neurons in the classification of a sequence. These are components, an input gateway, a neuron with a self-recurrent connection, a forget gate and an output gate. The self-recurrent connection ensures that all interference from the outside is ignored, allowing the state of a memory cell to be kept constantly between one time period and another. The input-gate controls the interactions between the memory cells and the environment. The input gate may block or allow input signals to change the state of the memory cell. The output gate can allow or prevent the state of a memory cell from having an effect on other neurons. Finally, the forget gate controls the recursive retro connection of the memory cell, which allows the cell to remember or forget its previous stage when needed.

2.8. Concluding Remarks

Regarding the recognition of characters, in general terms, the state-of-the-art (or, in other words, the best results over benchmark data sets) is basically composed by two main strategies, or based on these models: deep learning (Section “Deep Learning”) and committees of classifiers (Section “Committees”), which may be formed by deep models or convolutional networks. More details about preprocessing and classifiers applied to handwriting are presented in the next Chapters. At this point, we bound the discussion over the models, letting the general discussion about the results themselves and possible improvements to later sections.

Both leading general approaches to handwriting recognition try to overcome well-known difficulties of traditional methods or, in other words, to increase their accuracy and generalization power. And, moreover, a great part of core concepts of fundamental ideas

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 24 of 404

remain the same, the difference being the organization of the machines, either inside the model itself (in the case of deep learning), or externally (if we consider the ensembles). That fact comes out to be logical, although a bit controversial, when we remember that the neurons or basic processing units have not changed, and the representation and distribution of the knowledge over the network have been altered in deep learning, but not in its intrinsic concept: values and weights run forward and backward to adjust the model. This preamble does not obscure the virtue of recent methods or is intended to do so. It is only meant to make this bridge and pay off credits to their predecessors.

Indeed, there are several improvements and interesting ideas in these methods. Deep learning has the appealing motivation of removing feature engineering, which is one of the greatest difficulties when working with neural networks. It is known that if we have good features, almost all classifiers could discriminate these data. However, to encounter the best set of features or at least a good one may be a hard task. Even because, although humans are very proficient in reading, not necessarily we know what characteristics our brain uses to recognize the numbers, for example. Thus, we could only imagine what features are relevant and discriminant. And the performance of learning algorithms, in this case, depends on the quality of features too. Therefore, in deep learning, the features themselves are learned and coded inside the model itself and, consequently, the designer of the network does not need to concern about that. Of course, network architecture remains a function of the designer work, although it is claimed that the impact of architecture is somewhat reduced in deep models, or in other words, it is not expected to have drastic variations due to small modifications in the network structure.

On the other path, in ensembles of classifiers, the idea is that several experts may propose better solutions to a given problem than only one of them. It is not hard to think if we have more specialized agents, each one covering some area, region or subject, mainly when the problem is too complicated, they will tend to cover, more easily, the whole spectrum of that matter. In addition to this, if we also have an efficient mechanism to merge or select the best answer(s), more precise outcomes tend to be achieved. Thus, since each expert knows its sub-area and therefore provides meaningful or suboptimal answers regarding her localized knowledge, a “conference” of an experts group brings out the response for the given input, which is the global optimal answer in the best scenario.

3. Applications

3.1. Digits

There are various proposals in the literature to the specific task of recognizing handwritten Hindu-Arabic digits only. That especialization comes in general to simplify the problem, since, in a broader sense, digits may be regarded as characters. At first, we reduce the number of classes to ten (digits varies from zero to nine), and consequently there are fewer confusion possibilities, because only some pairs of digits are intrinsically similar. Moreover, digits natural intra-class variability is smaller than that of characters. In addition to those implementation issues, there are real world problems in which only digits need to be handled, such as automatic postal addressing, the processing of courtesy amount in bank checks, or the processing of dates or page numbers in documents to provide automatic

Cerence Operating Company Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 25 of 404

search and indexing. Thus, when one focuses on applications such as those, it makes sense to adopt recognition strategies specially designed to work on digits.

At this point, we need to resemble that these modifications are not always related to the core of the method. For example, neural networks are used to classify numbers, characters, or both at the same time, and the algorithm per se is not modified. However, different architectures may be used to deal with each situation. That also holds true if a principle such Occam's razor is considered (i.e., if we may use a simpler method, generally it is the best solution). So, if a simpler architecture may treat digits when working over digits, it is not necessary to use a more complex solution designed to handle characters in general.

Other question regarding digits as characters for recognition is the string length. Most of the papers in the literature focus on isolated digits. In other words, the digits need to be segmented. Segmentation is one of the most complicated tasks in document processing and is another prolific research area, and due to the scope of this chapter, we do not address that issue. However, there are methods that are applied to a numeric string as a whole, without performing a segmentation step, analogous to the case of word recognition. In Chapter 3, some issues of document processing and image segmentation are investigated and the principal literature works are addressed.

However, as we illustrate in Sections "Deep Learning" and "Committee", these are relatively complex models, which require more samples, training, computational cost, and efforts. As we show in Section "Digits", other simpler methods can achieve interesting results, whilst not as accurate, but with much fewer efforts in training and classification itself.

3.2. Characters

There are several factors that make character recognition a more complex task than digit recognition. In this scenario, one takes into account the existence of a greater number of classes to be recognized (this set can be composed of digits, uppercase, lowercase or accented characters, punctuation and symbols). The variation of the calligraphy of the individuals during the act of writing, the high similarity between distinct characters, as well as the change of the style of writing over time, are characteristics that also make the process of recognition of handwritten characters a complex activity. It is, therefore, noticeable that there is a high level of variability of the instances that can be correctly assigned to the same class. On the other hand, the high similarity between some distinct characters also increases the occurrence of false positives.

The task still becomes more challenging, as it is usually specific to each domain and application. In this way, the techniques established and used for the recognition of a certain family of characters, such as Latin / Roman, can not be applied in the same way for characters of different origin, such as Indus, Chinese or Arab.

That variability in writing makes the field of research related to the recognition of handwritten characters extensive so that the number of problems that must be addressed in order to obtain a satisfactory result in those various scenarios is enormous. That fact contributes to the existence of several ramifications of research in the area. In today's academic and industry environment, research can be found that works on the recognition of the various stages related to a recognition system.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

Among those stages, the final phase of character classification received a significant gain with the introduction of Deep Learning techniques. Today, in the offline character recognition task, the main approaches use deep multidimensional networks and deep neural network committees.

3.3. Words or Sentences

Word recognition refers to the process of segmenting the word regions of a text line and recognize it considering the whole character string. In sentence recognition, the classifier assumes text lines as input and normally the segmentation process is only needed in word spotting methods. Traditional modeling approaches based on Hidden Markov optical character models (HMM) and an N-gram language model (LM), as well as new approaches based on Multi-directional Long Short-Term Memory Neural Network (MDLSTM NN) have been used. In Chapter 3, different approaches of segmentation and recognition of words or sentences are investigated.

4. Selected Works

In this section, we introduce some of the most relevant papers in handwriting recognition. At this scenario, we consider works with the focus on digits, characters, or even both. Some papers still cover other applications, although they are out of the scope of this chapter. Because of this, we decided to not dive into separated subsections for each regarded subarea (as evidenced in Sections “Applications” and “Results”). The paper selection was guided by their reported accuracy on benchmark databases, which is an interesting criterion since this measure tends to be a good indicator of the merit of techniques.

4.1. Decoste and Schölkopf (2002)

Decoste and Schölkopf (2002) defined a new method for training Support-Vector Machines, that takes into account prior knowledge about the invariances of a classification task. In doing so, they reported high accuracy results in digit recognition, and also reduced training time when compared to other SVM-based methods.

By prior knowledge, we can understand as information about the learning task which is available in addition to training examples. In the most general sense, this knowledge is what makes it possible to generalize from the training samples to novel test examples (Decoste and Schölkopf, 2002). The paper deals with one specific type of prior knowledge: invariances. For instance, in image classification, there are transformations which do not change class membership (e.g. translations).

According to Decoste and Schölkopf (2002), there are three strategies to incorporate invariances into SVMs: (i) engineer kernel functions which lead to invariant SVMs; (ii) generate artificially transformed examples from the training set, or subsets thereof (e.g. support vectors), named as virtual examples; (iii) combine the two approaches by making the transformation of examples part of kernel definition, defined as kernel jittering. The two later approaches were focused in their paper. We should mention that virtual examples make

Document Ocr Text Corpus - Exhibit 2014
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 27 of 404

training time larger. Nevertheless, the authors try to diminish that influence by employing heuristics which make training more efficient, even with the inclusion of invariances.

At this point, to demonstrate the potential of using virtual examples (see Figure 8), consider we have prior knowledge indicating that the decision function should be invariant to horizontal translations. Therefore, the true decision boundary is given by the dotted line in the first frame (top left). However, due to different training examples, different separation hyperplanes are fully possible (top right). SVM would calculate the optimal hyperplane, as shown in Section 2.3 (bottom left), which is very different from the true boundary. In that case, the ambiguous point, denoted by the question mark, would be wrongly classified. The use of prior knowledge and consequent generation of virtual support vectors (VSV) yields a much more accurate decision boundary (bottom right), and leads to the correct classification of the ambiguous point.

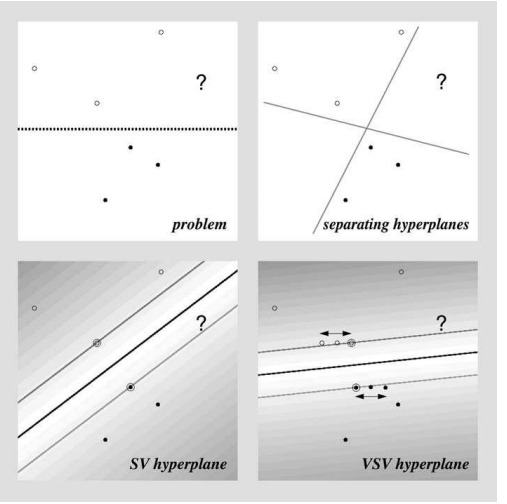


Figure 8. More accurate decision boundary by virtual support vectors (from (Decoste and Schölkopf, 2002)).

Specifically, they have developed two main heuristics, related to each other, over the SMO (Sequential Minimal Optimization) algorithm presented by Platt (1999) (although implementation and tests were conducted over an enhanced version of SMO, described by Keerthi et al. (2001)): (i) maximization of cache reuse, and (ii) *digestion*, the reduction of intermediate SV bulge. Cache reuse is important because most of the time spent at training dues to kernel matrix calculations. Besides, it is common that the same value is required many times. Therefore, if those values are stored in a "cache", redundant calculation time is saved. Digestion takes place when no additional support vectors are able to cache their kernel computations or, in other words, when the number of support vector set exceeded cache size. That issue is more severe in the case of virtual examples, since much more data is generated (intermediate SV bulge). The basic idea is to jump out of full SMO iterations early, once the working candidate support vector set grows by a large amount. Digestion allows for better control on the intermediate SV bulge size, besides enabling a trade-off between the cost of overflowing the kernel cache and the cost of doing as many inbound iterations as the standard SMO would.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 28 of 404

In their work, Descoste and Schölkopf performed a series of experiments in order to achieve high accuracy results on digit recognition, regarding the SVM training methodology. Their best result on MNIST, which is still state-of-the-art when one considers the use of Support-Vector Machines, was obtained employing the following settings: deslanted images, polynomial kernel of degree 9; $C = 2$, $3x3$ box jitter for VSV, consisting of translation of each image by a distance of one pixel in any of the eight directions (horizontal, vertical or both) combined with four additional translations by two pixels (horizontally or vertically, but not both).

In that case, the number of training examples was increased about 50%, while training time was increased four times when compared to the approach without additional translations; however, recognition rate was significantly improved. The number of VSVs, although apparently large (23,003 in the worst case), is still only about a third of the size of the full training set (60,000), despite a large number of explored translation jitters.

The authors illustrate an interesting approach for improving SVM training. It is important to observe that were proposed modifications in training itself, and also modifications on the data (by the construction of virtual samples). Those improvements made results much better than its predecessors and training time was reduced. Yet, other distortions, such as rotation, scale or line thickness, may be experimented to achieve even better and more general results.

4.2. Keysers et al. (2007)

Keysers et al. (2007) introduced a method for handwritten digit recognition based on image matching. Classification by flexible matching of an image is considered a fair approach to achieve low error rates. The images are distorted or transformed according to some different nonlinear deformation models, and recognition results are greatly improved, even using a simple classifier as k-NN. Those models are especially suited for local changes as they often occur in the presence of image object variability.

We can understand deformation of an image as the application of a two-dimensional transformation of the image plane, e.g., a small rotation or shift of a small part of the image. Matching of two images consists of finding the optimal deformation from a set of allowed deformations, in the sense that it results in the smallest distance between the deformed reference image and the observed test image. In addition to those, an important concept is the context of a pixel, which refers to the values of pixels in a neighborhood of that pixel and quantities derived from those.

There are several distinct deformation models with varied complexities. Basically, there exist zeroth, first, and second-order models (crescent degree of complexity). Experiments were made in order to testify the necessity or not of more complex models, with fewer constraints in contrast to simpler methods which have more constraints. In addition to this, the authors wanted to know if using pixel context information, i.e., pixels neighborhood values, would also improve the performance of the models on that recognition task.

In order to present an idea about the effects of applying this kind of deformation models, Figure 9 illustrates examples of nonlinear matching over handwritten digits. The first column shows the test and reference image. The rest of upper row exhibits transformed reference images using the indicated models, which best match with the test image. The lower

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 29 of 404

row shows the respective displacement grids generated to obtain the transformed images. The first two rows present results from digits belonging to different classes, while the two later, for digits from the same class. The examples on the left consider only image gray values. That one on the right shows the results using local context for matching (by the first derivative, obtained via Sobel filtering).

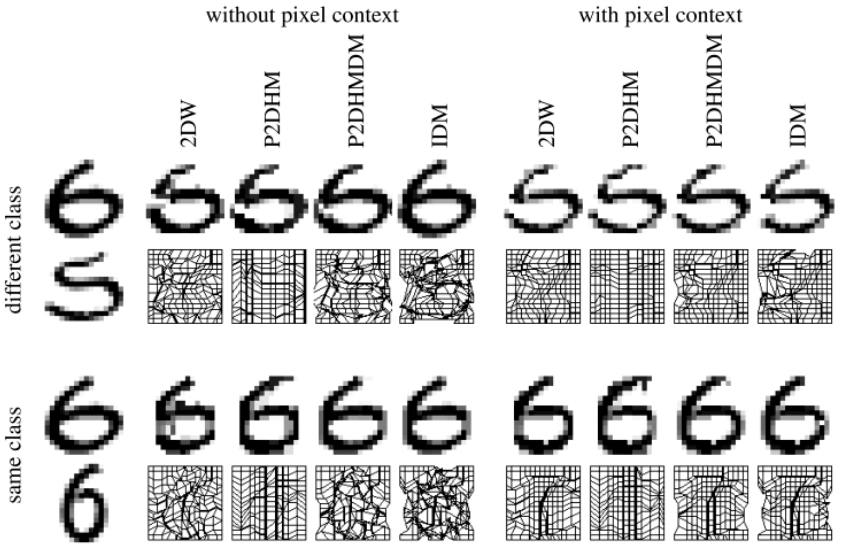


Figure 9. Nonlinear matching applied to digits images (after (Keysers et al., 2007)).

It is possible to observe that, for the same class, the models with more restrictions produce inferior matches to the models with less restrictive matching constraints (please concentrate on the left side). In the case of local context (right side), notice that the matchings for the same class remain very accurate, while the matching for the different classes is visually not as good as before (especially for models with fewer constraints, such as the IDM). Note also that the displacement grid is more homogeneous for matchings of the same class.

Thus, using this kind of artificially generated examples led to better results on handwritten digits recognition. Specifically, in conjunction with local context, the deformation model which obtained the best result was P2DHMDM (Keysers et al., 2004a)(Keysers et al., 2004b) (standing for Pseudo-two-dimensional hidden Markov distortion model). Other deformation model using pixel local context which achieved competitive results was IDM, which is a simpler model, allowing a trade-off between complexity and accuracy.

4.3. Cireşan, Meier and Schmidhuber (2010)

The work of Cireşan et al. (2010) is an excellent attempt to use a simple model, such as MLP with backpropagation, in contrast to the increasing complexity models found in the literature. Despite of its simplicity, their model achieved very accurate recognition rates. The main novelty came from using an MLP with several layers and many neurons per layer, thus opposing to the majority of related literature dealing with recognition of handwritten digits. The motivation was the following questions (Cireşan et al., 2010): Are all these

Cireşan Operating Company, Ex. 2114
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 30 of 404

complications of MLP really necessary? Can't one simply train really big plain MLP on MNIST?

Initial thinking may indicate that deep MLP does not seem to work better than shallow networks (Bengio et al., 2007). Training them is hard because backpropagated gradients vanish exponentially in the number of layers (Hochreiter et al., 2001). A serious problem affecting the training of big MLPs was processing power; training that kind of structure is unfeasible when considering conventional CPUs. Because of that, Cireşan, Meier and Schmidhuber (Cireşan et al., 2010) also make use of graphical units (GPUs), which permit fine-grained parallelism. In addition to that, the network is trained on slightly deformed images, continually generated online, i.e., created in each iteration; hence, the whole undeformed training set is available to validation, without wasting training images.

Detailing used strategies, training is performed using standard online backpropagation (Russel and Norvig, 2010), without momentum, but with a variable learning rate. Weights are initialized with a uniform random distribution; and activation function of each neuron is a scaled hyperbolic tangent (after (Lecun et al., 1998)). The images deformations were: elastic distortions (Simard et al., 2003); an angle for either rotation or horizontal shearing; and horizontal and vertical scaling.

Some MLP architectures were investigated (Cireşan et al., 2010). The one which yielded the best results was: 784, 2500, 2000, 1500, 1000, 500, 10, each number meaning the number of neurons of the layers, being the first one the input layer (784 neurons because the input images are 28x28), and the last one, the output layer (that of course presents 10 neurons, since the problem at hand is digit recognition); totalizing 12.11 million weights. Other interesting information about the training procedure were the outcomes obtained by the use of GPU: deformation routine was accelerated by a factor of 10; forward and backward propagation were sped up by a factor of 40.

The performed experiments proved simple plain deep MLP can be trained. Even the huge number of weights could be optimized with gradient descent, achieving test errors below 1% after 20 to 30 epochs in less than two hours of training. In part, the explanation comes from the continual deformations of the training set, that generate a virtually infinite supply of training examples, and the network rarely sees any training image twice or indefinitely, what seems to be the case in normal backpropagation training, causing saturation to the network.

4.4. Cireşan et al. (2011)

Cireşan et al. (2011a) introduced a convolutional neural network committee for handwritten character classification. The motivation was two-fold: (i) CNN are among the most suitable architectures for character recognition; and, (ii) the sets of misclassified patterns by different classifiers do not necessarily greatly overlap. Thus, it would be possible to improve recognition rates if the errors of classifiers on various parts of the training set differ as much as possible. They try to achieve this by training identical classifiers on data pre-processed or normalized in different ways (Cireşan et al., 2011b).

The same architecture was used for both digits and characters experiments. Nets have an input layer of 29x29 neurons, followed by a convolution layer with 20 maps of 26x26 neurons, and 4x4 filters. After, a max-pooling layer with a 2x2 kernel whose outputs are

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 31 of 404

connected with another convolution layer containing 40 maps of 9x9 neurons each. The last max-pooling layer reduces map size to 3x3, using 3x3 filters. A fully connected layer of 150 neurons is connected to the max-pooling layer. Output layer has one unit per class, and therefore, they have 62 neurons for characters and 10 for digits. All CNNs are trained in a full online mode with annealed learning rate and continually deformed data (elastic deformation, rotation and horizontal and vertical scaling, as made in (Cireşan et al., 2010)). Also, GPUs were used to accelerate all training procedure.

Experiments were performed on the original and six preprocessed data sets. Preprocessing was motivated by different aspect ratios of characters caused by writing styles variations. The width of all characters were normalized to 10, 12, 14, 16, 18, 20 pixels, except for characters "1", "i", "I" and "l", and the original data (Cireşan et al., 2011b). Figure 10 illustrates training and testing strategy. Training is shown in item a: each network is trained separately and normalization is done prior to training. During each training epoch, every character is distorted in a different way, and the data is fed to the network. The committees are formed by averaging corresponding outputs (item b).

For each of the datasets (original or normalized), five CNNs with different initialization are trained for the same number of epochs (resulting in a committee formed by 35 CNNs). Consequently, it is possible to analyze output errors for the $5^7 = 78125$ possible committees of five nets, each trained on one of the seven data sets.

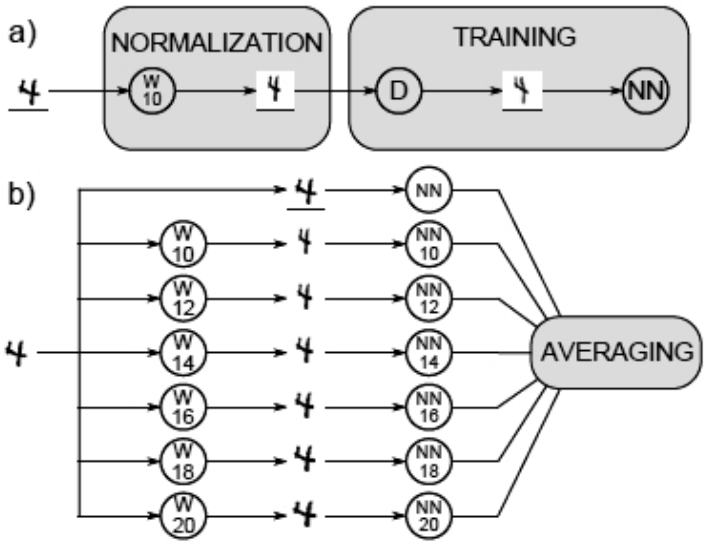


Figure 10. Classification strategy, (a) training a committee member, (b) testing with a committee (from (Cireşan et al., 2011a)).

Therefore, simple training data preprocessing led to experts with less correlated errors than those of different nets trained on the same bootstrapped data. Thus, simply averaging experts outputs considerably improved recognition rates. It was credited the first time automatic recognition really comes near to human performance (Lecun et al., 1995)(Kimura et al., 1997).

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

4.5. Cireşan, Meier and Schmidhuber (2012)

Cireşan et al. (2012) proposed a multi-column approach for Deep Neural Networks (DNN) based on small receptive fields of convolutional winner-take-all neurons yield large network depth, resulting in roughly as many sparsely connected neural layers. Only winner neurons are trained. The authors suggest the several deep neural columns become experts on inputs preprocessed in different ways and the result is the average of their predictions.

The proposed architecture and its training and testing procedures are illustrated in Figure 11.

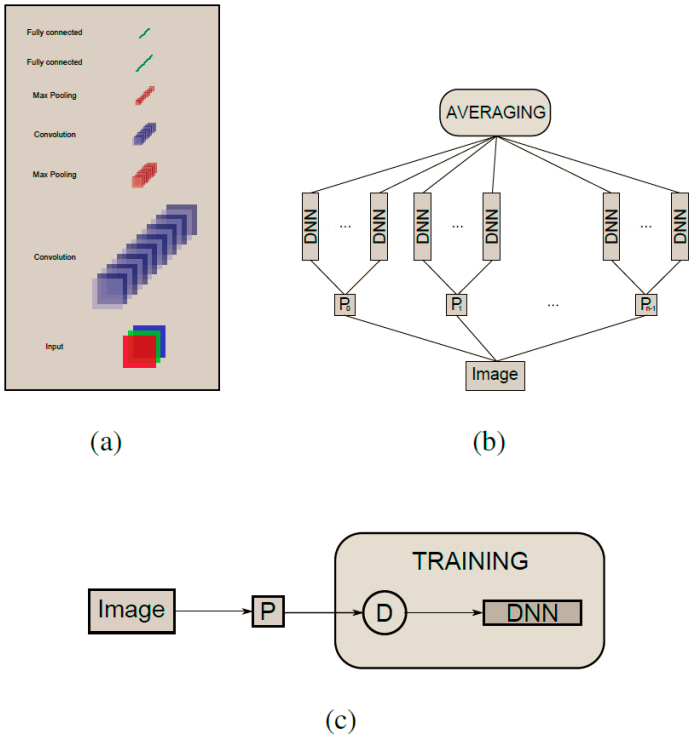


Figure 11. (a) DNN architecture. (b) MCDNN architecture. The input image can be preprocessed by $P_0 - P_{n-1}$ blocks. An arbitrary number of columns can be trained on inputs preprocessed in different ways. The final predictions are obtained by averaging individual predictions of each DNN. (c) Training a DNN (from Cireşan et al. (2012)).

The authors combine several techniques to iteratively train DNN in a supervised way. They use hundreds of maps per layer, with many (6-10) layers of non-linear neurons stacked on top of each other. The overlapped receptive share weights of 2-dimensional layers uses winner-take-all. Given some input pattern, a max pooling technique determines the winning neurons. It allows the selection of the most active neuron of each region. The winners of some layer represent a smaller, down-sampled layer with lower resolution, feeding the next layer in the hierarchy (Cireşan et al., 2012). The receptive fields and winner-take-all use 2x2 or 3x3 neurons. The DNN columns are combined to form a Multi-column DNN (MCDNN). Given some input pattern, the predictions of all columns are averaged.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 33 of 404

$$y^i MCDNN = \frac{1}{N} \sum_{j}^{\#columns} y^i DNN_j \quad (1)$$

The experiments are performed with MNIST, NIST SD 19, Chinese characters, NORB, Traffic signs and CIFAR 10 datasets. The authors claim that is the first time human-competitive results were achieved on widely used computer vision benchmarks. As it can be seen in Table 1, the obtained results are impressive. On many image classification datasets, the MCDNN improves the state-of-the-art by 26-80%.

Table 1. Comparison of MCDNN and literature approaches on different datasets (from Cireşan et al. (2012)).

Dataset	Best results on literature (details on Cireşan et al. (2012)) [%]	MCDNN [%]	Relative improvement [%]
MNIST	0.39	0.23	41
NIST SD 19	30.91*	21.01*	30-80
HWDB1.0 on.	7.61	5.61	26
HWDB1.0 off.	10.01	6.5	35
CIFAR10	18.50	11.21	39
traffic signs	1.69	0.54	72
NORB	5.00	2.70	46

* Letters with 52 classes. For global results see Cireşan et al. (2012) on Table 4.

4.6. Yuan et al (2012)

Yuan et al. (2012) applies Convolutional Neural Networks for offline handwritten English character recognition. They use a modified LeNet-5 CNN model, with special settings of the number of neurons in each layer and the connecting way between some layers. Outputs of the CNN are set with error-correcting codes, thus the CNN has the ability to reject recognition results. For the CNN training, an error-samples-based reinforcement learning strategy is developed.

The CNN model used in this paper is a modified LeNet-5. Many modifications are made in the basic architecture of LeNet-5, to attain a tradeoff between time-cost and recognition performance. First, they change the number of neurons in each layer of a CNN, thus different models are made. In this models, a new symmetrical connection is made to overcome the loss of parameters and information caused by the asymmetric connection between two convolutional layers on original LeNet-5. That new connection produces a symmetrical map, each feature map in the posterior layer connects to more feature maps in the predecessor layer, considering the redundancy of features and the time cost.

In addition to the improvements surrounding the structure of CNN. The article proposes a new method to improve the training stage of convolutional networks. The Error-Samples-Reinforcement-Learning algorithm (ESRL) method uses the instances erroneously classified during the training step and from there they generate a new set of those modified images with preprocessing, so that they can be used in later training stages together with some images that were well classified. The technique seeks to decrease the network error more quickly, reinforcing training with variations of examples in which it obtained a considerable error rate.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 34 of 404

Experiments are evaluated on UNIPEN lowercase and uppercase datasets, with recognition rates of 93.7% for uppercase and 90.2% for lowercase, respectively. All uppercase or lowercase samples are randomly divided into 3 subsets. Training is on the first 2 subsets; 33% and 67% of the 3rd subset are validation set and test set respectively. Training is repeated 3 times.

5. Literature Results

5.1. Digits

Here, we present some of the most recent and relevant results on single digit recognition (summarized in Table 2). For each method exposed in this table, we indicate the underlying model, according to what is explained in Section 2, the used features or the absence of them (denoted by N/A, standing for not applicable), and the error rate (in percentage) over the MNIST database.

Table 2. State-of-the-art results digits recognition

Underlying models	Method	Features	Error rate
Conv., deep, committees	MCDNN (Cireşan et al., 2012)	N/A	0.23
	CNN Committee (Cireşan et al., 2011a)	N/A	0.27
Deep, MLP	Deep, big, simple MLP (Cireşan et al., 2010)	N/A	0.35
kNN	kNN / nonlinear deformation (Keysers et al., 2007)	N/A	0.52
SVM	Virtual SVM / jitter (Decoste and Schölkopf, 2002)	N/A	0.56

Table 2 shows that there are several interesting results for digits recognition. In this context, merit is twofold: (i) error rates are very low, and reached that of human performance (Cireşan et al., 2012); and (ii) there is not only one kind of model that reaches good accuracy. As the main objective, the accuracy of a recognizer is its fundamental reason, and therefore, those rates reaffirm there are methods able to handle handwritten digits to a certain extent (of course, this is a reference database, and different problems and tricky examples appear all the time). Different applications, or even practical issues of implementation or of model training, and even about the technical team skills, may favor one technique over another. For instance, if one does not possess suitable hardware, use of deep learning is unfeasible (graphical cards are still expensive today). Moreover, when the hardware is proper, it is still needed that people have specific programming skills, such as parallel programming, CUDA, and other software libraries. In such a situation, the use of kNN or SVM-based methods may be an appropriate choice if a possible accuracy loss may be tolerated. Another factor may be the "time to market", since a simpler model may be implemented quickly, and depending on deadline requirements, this technique may appear to be more adequate.

In addition to this, one may think the results shown above indicate that digit recognition is solved. However, we need to remind this scenario considers databases with "c"an" digits, or in other words, those images have no noise or another kind of artifacts that could impair the classification process itself. Another aspect relates to the data distribution on the datasets. In MNIST, for instance, although the original configuration states 60,000 test sam-

Confidential Document - Subject to Protective Order
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 35 of 404

ples, the common practice is to use a subset of 10,000 samples. At first sight, this number appears to be significant. But, if we remember the huge quantity of processed images and data daily, and consequently, the uncountable number of digits to be treated, this amount may be insufficient to a practical or business point of view (of course, this fact does not invalidate the findings of the area over this database). The use of other datasets as the NIST SD19 is welcome but lacks standardization, since it does not provide a default or suggested a division of data between training and test. That may cause misinterpretation or confusion of results because the algorithms are not necessarily evaluated using the same data partition.

Therefore, we think new databases are needed, mainly when considering the required quantity of data for training of deep learning models. In this context, the number of examples should be increased, as well the variations over the data sets. It is also latent the necessity of scrutinized evaluation of the methods, since many papers do not run statistical tests and, in various cases, the difference between them is minimal. Of course, this fact does not disregard the proposition of new techniques and, obviously, we are not stating new methods are only valuable when overcome the performance of others in terms of error rate, since there are several others factors about the merits of published works, such as computational cost, speed of convergence, simplicity or complexity of the models, etc.

5.2. Characters

In this section, we present some of the most recent and relevant results on character recognition (summarized in Table 3). For each method exposed in this table, we indicate the underlying model, according to what is explained in Section 2, the dataset used in the experiments, and the error rate in percentage for classification the uppercase and lowercase letters.

Table 3. State-of-the-art results characters recognition

Underlying models	Method	Dataset	Error rate (uppercase-lowercase)
Conv., deep, committees	MCDNN (Cireşan et al., 2012)	NIST-19	1.83 - 7.47
Conv., deep, committee	CNN Committee (Cireşan et al., 2011a)	NIST-19	1.91±0.06 - 7.71±0.14
Conv.	CNN (Yuan et al., 2012)	UNIPEN-HECR	6.3 - 9.8

We can see in the Table 3 that the best methods for classification of isolated characters are the manuscripts based on convolutional networks. The convolutional networks committees as well as the task of digit classification obtained by far the best results, even with the greatest difficulty of the task addressed. The biggest cause of error in the classification is the ambiguity between some characters like Í, ítha t are impossible to be treated if context information is not taken into account.

However, it is worth noting that the problem of character classification has not yet been completely solved since the good results were obtained in very specific subgroups of data. In the classification of handwritten characters, the results are better when the networks are trained and tested in bases with specific characteristics, where the images have the same size, width and variation pattern. However, when a set of tests with a greater variation and number of classes is used for validation of the method, the results are somewhat lower. It

Cerence Operating Company, IPR2024-01267
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 36 of 404

is soon observed that the greater the number of classes introduced in the problem at hand, the higher the chance of ambiguity and error in the classification. We can note this in (Cireşan et al., 2012) and (Cireşan et al., 2011a) when using a set consisting of upper and lower characters to validate the results are less than the results using the sets individually as shown above. The results could then be quite different with different databases.

Observing those problems, we note that there are new challenges that must be addressed, since in the results there were no accented characters, symbols and punctuation. In other alphabets, the amount of characters is much greater than those belonging to the Latin alphabet so, in those cases, new strategies must be elaborated. It has been found that the number of papers in the classification of Latin manuscripts is much larger than the number of papers involving other families of handwritten characters. Thus, there is more need of research in this area, including the tackling of the two problems mentioned above (which although are similar, have some differences).

6. Trends and Ideas

From the analysis and results obtained by the different methods and models reported in the literature (Sections "Models" and "Selected Works"), it is possible to glimpse main tendencies and extrapolate for some possible future directions. Regarding about current trends, we observe a predominance of deep and convolutional models to recognize characters in general (and other objects too, although this discussion is out of scope). Another strong trend in handwriting recognition is the use of deformation or distortion techniques prior to training. There seems to be a consensus about the generation of artificial examples simulating deformation or distortions which may reflect natural variations of handwriting and consequently, these preprocessing strategies may yield better recognition results.

Until now, those models proved to be very accurate, although there are some limitations, such as the huge number of samples required for their learning, or the continuing lack of comprehension about the obtained results (the latter being a problem already observed on shallow networks, such as multilayer perceptrons). Even though that problem is around for a long time, no significant evolution has been noticed on it over the past years. About the accuracy, one may argue: are models actually learning more or they are just overfitting on training data? Other question could be: how much improvement was really achieved by those newer models in contrast to older ones? As we may observe in 2, traditional methods are not so far from the newest ones in the context of digits recognition.

Another issue regarding those more complex models is the training time; it is expected more robust and precise methods need more time and computational efforts, but in some situations the use of that kind of technique may be prohibitive due to cost or time constraints. Even forward propagation in deep networks demands a great amount of time. So, in this context, we may ask: would it be possible to accelerate deep neural networks training? Would it be possible to have the same learning capacity with less complex architectures (thus reducing the number of nodes and, consequently, the time required for their training)?

Moreover, the evaluation of deep networks seems to be somewhat lacking. We understand time is a hard constraint, and thus, the repetition of training and testing in rounds is difficult. However, we think that is unavoidable. This is true even when we resemble those networks are less sensitive to weights or parameters initialization (less sensitive does not

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 37 of 404

mean insensitive, which is the situation where repetition could be avoided).

Regarding evaluation and architecture in conjunction, there is no evidence about the minimum requirements for solving recognition tasks, including the case of character recognition. The papers do not explore, for instance, the accuracy of more compact networks, letting researchers choose between accuracy or processing time (including training and testing). Or in other words, no considerations are made about different architectures, which would be useful even if they achieved bad results, to show the chosen model is the best option, or to present simpler architectures with acceptable results (and let researchers pick the best fit for their scenario considering the time/precision/resources trade-off). That knowledge would certainly benefit the area of handwriting recognition as a whole.

Extrapolating recognition and thinking about reading at all, other interesting issues are related to the human ability to read, and in this sense, the incorporation of our "theoretical strategies" to read in several aspects. For instance, when we are faced with an ambiguous character, besides trying to recognize this symbol per itself, which is the basic operating mechanism of all recognition methods, we also exclude those symbols which do not apply to that image. For the sake of clarification, consider the following example (we are exposing a digit situation, but the idea may be easily applied or extended to characters in general): if we have a numeral "eight" that has its upper part degraded, present techniques will tend to confound it with other numbers since its constituent strokes are not clear. However, in the case of humans, besides analyzing the image thinking about what digit is written there, we also think that digit could not be a number "2" or "3" because of the closed loop at the bottom part of the number, eventually leading to the correct reading of the digit. By the best of authors knowledge, present methods do not possess that kind of complementarity, and we understand and believe that sort of reasoning may help to advance the state-of-the-art in recognition of handwritten characters.

Another point, even in the case of isolated character recognition, is the use of context. At first sight, one may think this idea only applies when dealing with words or sentences. However, we humans use context, even when we do not have the notion about all the situation in which we are posed. And, we know the more we know about the context, the more precise and accurate we tend to be at interpreting that scenario. Yet, usually, when thinking about the context in character recognition, the major part of strategies consist of post-correction of characters. For instance, if we know that a field is a date, we may correct some digits or characters based on the date constraints. Nevertheless, it would be very different if that knowledge, to some extent, was introduced in the model itself. Thus, instead of correcting after recognition based on predetermined rules, the model could interpret what is being read, and from that, try to make a more precise reading of that information.

Thinking in a broader sense, character recognition is evaluated over isolated characters, with no idea of what is the read information in an up level analysis – in other words, which information is being read when the recognition of that character takes place? Is it a name, a date, a common word? We understand that reality, and also agree that in the context of algorithm evaluation, that scenario does make sense. However, we guess that the fact of recognizing isolated symbols does not prevent the use of prior knowledge concepts also being applied on the information being recognized. That statement holds because, in the most part of the cases, we segment characters and feed them to the classifier aiming to read the content of the whole document. Thus, the recognition is made over a single character,

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 38 of 404

but we are in fact interested in recognizing all characters, which represent the document content.

Thus, we think those questions need to be considered when proposing new models or the design of new methods using existent models.

Acknowledgments

The authors acknowledge Document Solutions for sponsoring this research. The authors also thank CNPQ for supporting the project under the grant "Bolsa de Produtividade DT" (Process 311338/2015–1).

References

Bengio, Y., Pascal, L., Popovici, D., and Larochelle, H. (2007). Greedy layer-wise training of deep networks. In Schölkopf, P. B., Platt, J. C., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 153–160. MIT Press, Cambridge, MA, USA.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Bortolozzi, F., Britto Jr, A. S., de Oliveira, L. E. S., and Morita, M. (2005). Recent advances in handwriting recognition. In Pal, U., Parui, S. K., and Chaudhuri, B. B., editors, *Document Analysis*, pages 1–30.

Braga, A. P., Carvalho, A. P. L. F., and Ludermir, T. B. (2007). *Redes Neurais Artificiais*. LTC, Rio de Janeiro, second edition.

Cireşan, D., Meier, U., and Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3642–3649, Providence.

Cireşan, D. C., Meier, U., Gambardella, L. M., and Schmidhuber, J. (2010). Deep, big, simple neural nets for handwritten digit recognition. *Neural Computation*, 22(12):3207–3220.

Cireşan, D. C., Meier, U., Gambardella, L. M., and Schmidhuber, J. (2011a). Convolutional neural network committees for handwritten character classification. In *International Conference on Document Analysis and Recognition*, pages 1135–1139, Beijing.

Cireşan, D. C., Meier, U., Gambardella, L. M., and Schmidhuber, J. (2011b). Handwritten digit recognition with a committee of deep neural nets on gpus. Technical Report IDSIA-03-11, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, Manno, Switzerland.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314.

Decoste, D. and Schölkopf, B. (2002). Training invariant support vector machines. *Machine Learning*, 46(1-3):161–190.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification*. Wiley-Interscience, New York, second edition.

Fahlman, S. E. (1988). An empirical study of learning speed in back-propagation networks. Technical Report CMU-CS-88-162, Carnegie Mellon University, Pittsburgh.

Freund, Y., Schapire, R. E., et al. (1996). Experiments with a new boosting algorithm. In *Icml*, volume 96, pages 148–156.

Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier neural networks. *Journal of Machine Learning Research*, 15(106):275.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge, MA, USA.

Hagan, M. T. and Menhaj, M. B. (1994). Training feedforward networks with the marquardt algorithm. *IEEE Transactions on Neural Networks*, 5(6):989–993.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. Springer Series in Statistics. Springer, New York, second edition.

Haykin, S. (2009). *Neural Networks and Learning Machines*. Prentice-Hall, Upper Saddle River, third edition.

Hochreiter, S., Bengio, Y., Frasconi, P., and Schmidhuber, J. (2001). Gradient flow in recurrent nets: The difficulty of learning long-term dependencies. In Kramer, S. C. and Kolen, J. F., editors, *A field guide to dynamical recurrent neural networks*. IEEE Press, Piscataway, NJ, USA.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Huang, F. J., Boureau, Y.-L., LeCun, Y., et al. (2007). Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE.

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., and Murthy, K. R. K. (2001). Improvements to platt's smo algorithm for svm classifier design. *Neural Computation*, 13(3):637–649.

Keysers, D., Deselaers, T., Gollan, C., and Ney, H. (2007). Deformation models for image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8):3207–3220.

Keysers, D., Gollan, C., and Ney, H. (2004a). Classification of medical images using non-linear distortion models. In Tolxdorff, T., Braun, J., Handels, H., Horsch, A., and Meinzer, H., editors, *Bildverarbeitung für die Medizin 2004: Algorithmen — Systeme — Anwendungen*, pages 366–370. Springer Berlin Heidelberg, Berlin.

Keysers, D., Gollan, C., and Ney, H. (2004b). Local context in non-linear deformation models for handwritten character recognition. In *17th International Conference on Pattern Recognition*, pages 511–514, Cambridge, UK.

Kimura, F., Kayahara, N., Miyake, Y., and Shridhar, M. (1997). Machine and human recognition of segmented characters from handwritten words. In *International Conference on Document Analysis and Recognition*, pages 866–869, Ulm, Germany.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

LeCun, Y. and Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lecun, Y., Jackel, L. D., Bottou, L., Cortes, C., Denker, J. S., Drucker, H., Guyon, I., Müller, U. A., Säckinger, E., P., S., and Vapnik, V. (1995). Learning algorithms for classification: A comparison on handwritten digit recognition. In Oh, J. H., Kwon, C., and Cho, S., editors, *Neural Networks: The Statistical Mechanics Perspective*, pages 261–276. World Scientific.

Mello, C. A. B., Olivera, A. L. I., and Santos, W. P. (2012). *Digital Document Analysis and Processing*. Nova Science Publishers, Inc., Commack, NY, USA.

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. In Schölkopf, B., Burges, C. J. C., and Smola, A. J., editors, *Advances in Kernel Methods*, pages 185–208. MIT Press, Cambridge, MA, USA.

Riedmiller, M. and Braun, H. (1993). A direct adaptive method for faster backpropagation learning: the rprop algorithm. In *IEEE International Conference on Neural Networks*, pages 586–591, San Francisco.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 41 of 404

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986a). Learning internal representations by error propagation. In Rumelhart, D. E., McClelland, J. L., and PDP Research Group, C., editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*, pages 318–362. MIT Press, Cambridge, MA, USA.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986b). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1988). Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1.

Russel, S. and Norvig, P. (2010). *Artificial Intelligence*. Prentice-Hall, Upper Saddle River, third edition.

Sammut, C. and Webb, G. I. (2011). *Encyclopedia of machine learning*. Springer Science & Business Media.

Simard, P. Y., Steinkraus, D., and Platt, J. C. (2003). Best practices for convolutional neural networks applied to visual document analysis. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition-Volume 2*, pages 958–, San Mateo, CA. IEEE Computer Society.

Vapnik, V. (2000). *The Nature of Statistical Learning*. Springer, New York, second edition.

Yuan, A., Bai, G., Jiao, L., and Liu, Y. (2012). Offline handwritten english character recognition based on convolutional neural network. In *Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on*, pages 125–129. IEEE.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

In: Handwriting: Recognition, Development and Analysis ISBN: 978-1-53611-937-4
Editors: Byron L. D. Bezerra et al. © 2017 Nova Science Publishers, Inc.

Chapter 2

THRESHOLDING

Edward Roe[1,*] ***and Carlos Alexandre Barros de Mello***[2,†]
[1]CESAR - Centro de Estudos Avançados do Recife,
Recife, Brazil
[2] Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil

1. Introduction

Thresholding can be seen as a classification problem where, usually, there are two classes (Mello, 2012). In particular, for document images, it is expected that a thresholding algorithm correctly classifies the ink (the foreground) and the paper (the background). If to the ink is attributed the black color and to the paper the white color, the result will be a bi-level image or a binarized image. This is why thresholding is also known as binarization. Considering, for example, a digital grayscale image, the process is quite simple: given a threshold value, says th, the colors above this value are converted into white, while the colors below are converted into black, separating both classes. The problem is to correctly find the threshold value that makes a perfect match for foreground and background elements. This is the major concern of thresholding algorithms. The problem is much more complex when we deal with natural scenes images (as the concept of foreground and background is not so clear). For document images, it is easier to understand what is the expected result although there are several issues that make this domain so difficult as aging degradations like foxing (the brownish spots that form on the paper surface), back-to-front ink interference, illumination artifacts (like shadows, due to acquisition process), crumpled paper and adhesive tape marks. Some of these problems can be seen in Figure 1

For document images, thresholding is useful and the first step for several processes as skew estimation (Brodić et al., 2014) and correction (Ávila and Lins, 2005), line segmentation and text extraction (Sánchez et al., 2011), word spotting (Almazán et al., 2014), etc.

*E-mail address: edward.roe@gmail.com.
†E-mail address: cabm@cin.ufpe.br.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy



Figure 1. Examples of many kinds of problems caused by ageing process as (top right and bottom right) foxing, (bottom left and middle right) back-to-front interference and (top left) caused by human manipulation as adhesive tapes and (bottom right) crumpled paper.

Even more, several algorithms for character recognition work with bi-level images (Mello, 2012). An incorrect thresholding can bring consequences to all of these further processes. This can be seen in Figure 2 with samples of correct and incorrect thresholding

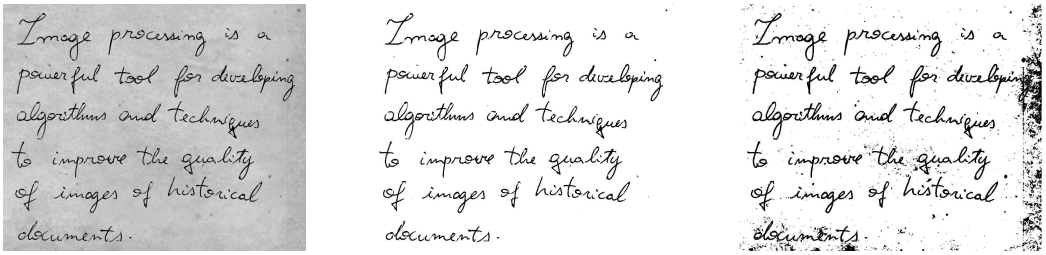


Figure 2. (left) Original document image in grayscale; (center) result after a correct separation of background and foreground; and (right) a threshold value to high converted many gray tones into black, misclassifying some of them, making it hard to read some words with no knowledge about the original document.

There are several different features that can be used to try to provide the correct separation between tones. Besides, a thresholding algorithm can be applied globally (a unique threshold value is used for the complete image) or locally (the image is divided into regions and each region has its own threshold value and even different algorithms). As examples of global algorithms we can cite Otsu, Pun, Kapur, Renyi, two peaks, and percentage of black; local algorithms can be exemplified by Sauvola, Niblack, White and Bernsen algorithms. For a general view of all these methods and many more, we suggest the reading of Sezgin and Sankur survey (Sezgin et al., 2004).

Entropy, statistical properties, stroke width estimation and histogram shape are just a few examples of common features that can be used to define the threshold value. In the rest of this chapter, we present more recent algorithms for the problem with more unusual approaches.

It is also important to observe that by document we mean any kind of paper with information stored. This generalizes the concept just to make easier the comprehension. Thus,

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

by document, we mean letters (both handwritten or typewritten), book pages, forms, topographic maps, floor plans, blueprints, sheet music, postcards, and bank checks. Of course, in most part of the chapter, we are going to present examples applied to letters or usual documents. Section “Graphical Documents” deals with other types of documents highlighting the major features that make them unique.

The next three sections present algorithms with different kind of approaches. As said before, Section “Graphical Documents” introduces the problem for unusual types of documents. In Section “Thresholding Evaluation”, we discuss the problem of automatic evaluation of binarization, followed by the conclusions of the chapter.

2. Edge Based Algorithms

The stroke edge as a strong text indicator has been used for document image thresholding (Sezgin et al., 2004). But for degraded document images, stroke edges may not be detected properly due to various types of document degradation.

The algorithm presented by Lu et al. (2010) makes use of both the document background and the text stroke edge information. It first estimates a document background surface through an iterative polynomial smoothing procedure. Then text stroke edges are detected based on the local image variation within the compensated document image. After the document text is extracted based on the local threshold, estimated from the detected text stroke edge pixels. Finally, some post-processing operations are performed to improve the binarization results. Each step is described next.

The document background surface is estimated through polynomial smoothing. The smoothing is done in three phases, first, the document background surface is estimated through one-dimensional polynomial smoothing that is usually faster and more accurate than the two-dimensional polynomial smoothing. Second, the global polynomial smoothing is performed, which fits a smoothing polynomial to the image pixels within each whole document row/column. The global smoothing polynomial is usually capable of tracking the image variation within the document background accurately considering that text documents usually have a background of the same color and texture. Third, after each round of smoothing, the polynomial smoothing is performed iteratively updating the polynomial order and the data points adaptively. The iterative smoothing further improves the accuracy of the estimated document background surface (Lu et al., 2010).

The text stroke edges are detected based on the local image variation. Before the evaluation of the local image variation, the global variation of the document image contrast is first compensated so that the text stroke edges can be better detected. The document contrast compensation is performed by using the estimated document background surface.

Lu, Su and Tan empirically observed that many edge pixels detected by the traditional edge detector do not correspond to the real text stroke edges within document images. Instead, the text stroke edge pixels can be better detected from the ones that have the maximum L1-norm image gradient in either horizontal or vertical direction. The local image variation at each candidate text stroke edge pixel is then evaluated by combining the L1-norm image gradient in both horizontal and vertical directions. First, a number of candidate text stroke edge pixels is detected by the ones that have the maximum L1-norm image gra-

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 45 of 404

dient in either horizontal or vertical direction, as defined by Equation (2.1).

$$V_h(x,y) = |\bar{I}(x,y+1) - \bar{I}(x,y-1)| \\ V_v(x,y) = |\bar{I}(x+1,y) - \bar{I}(x-1,y)| \tag{2.1}$$

where $\bar{I}$ is the normalized document image defined by Equation (2.2).

$$\bar{I} = \frac{C}{BG} \times I \tag{2.2}$$

where C is a constant that controls the brightness of the compensated document images and BG is the estimated document background surface.

The local image variation at each candidate text stroke edge pixel is then evaluated by combining the L1-norm image gradient in horizontal and vertical directions using Equation (2.3).

$$V(x,y) = V_h(x,y) + V_v(x,y) \tag{2.3}$$

The candidate text stroke edge pixels are detected being the ones having either the maximum $V_h(x,y)$ or the maximum $V_v(x,y)$.

The real text stroke edge pixels are detected by using Otsu's global thresholding method (Otsu, 1979) based on local image variation of the detected candidate stroke edge pixels, which histogram usually has a bimodal pattern.

Once the text stroke edges are detected, the document text can be extracted based on the observation that the document text is surrounded by text stroke edges and also has a lower intensity level compared with the detected stroke edge pixels.

Finally, three post-processing operations, based on the estimated document background surface and some document domain knowledge, are applied to correct the document thresholding error:

- Remove text components of a very small size that often result from image noise such as salt and pepper noise;
- Remove the falsely detected text components that have a relatively large size;
- Remove single-pixel holes, concavities, and convexities along the text stroke boundary.

The method proposed by Roe and Mello (2013) makes use of local image equalization and an extension to the standard difference of Gaussians edge detection operator, XDoG. The binarization is achieved after three main steps:

1. First binarization:
   A. Local image equalization
   B. Binarization using the Otsu algorithm
2. Second binarization:
   C. Global image equalization
   D. Edges detection using XDoG
3. Cleanup and restoration:

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 46 of 404

E. Combine the results from steps B and D
F. Remove noise from image generated in step E and Restoration, filling gaps

Each step is described next.

The main goal of local equalization is to prepare the image for the final binarization. The idea is to change the intensity differences between pixels, emphasizing it at opposite sides of a sharp edge and minimizing it for pixels in soft edges.

The local equalization, presented in Roe and Mello (2013), is performed through the following steps: first the image is converted into values in (0, 1] interval (0 value is converted into 0.01 to avoid division by zero) and it is scanned using a $n_{eq} \times n_{eq}$ window and the higher pixel intensity is found at each window. This intensity is used to divide the pixel intensity at the center of the window and the result is placed in a new image at the same position. As the degraded documents are yellowish/brownish, we get better results considering only the red channel of the image. To increase the contrast after the equalization, the gamma function $c = r^{\gamma}$ is applied to entire the image with $\gamma = 1.1$. Figure 3 shows the result of applying local equalization directly on sample images shown in Figure 1. The window size n_{eq} has impact on resulting edge thickness; larger windows result in thicker edges.



Figure 3. Example of local equalization algorithm applied to Figure 1.

Otsu binarization algorithm (Otsu, 1979) is used in this step to separate degenerated background regions from the text. In this process, some text could be also removed but this is not a great problem, as the idea here is just to use the result of the Otsu method as a guide in a further step. A cleanup is performed in Otsu's result to remove some remaining noise. This cleanup just removes contiguous areas having less than 10 pixels in size.

The Global image equalization step, applied to enhance image contrast before the XDoG filter, is similar to the first step but with two important differences: the equaliza-

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

tion is done directly over the entire image and using the three channels, red (r), green (g) and blue (b), independently. The results from each channel are then combined together.

The Difference of Gaussians (DoG) is an edge detector that involves the subtraction between two versions of the original grayscale image, blurred with a Gaussian filter, creating a band-pass filter which attenuates all frequencies that are not between the cutoff frequencies of the two Gaussians (Marr and Hildreth, 1980; Gonzalez and Woods, 2007). Better results were achieved using an extension of DoG, called XDoG (Winnemöller, 2011) given by Equation (2.4).

$$XDoG(\sigma, k, \tau) = G(\sigma) - \tau \times G(k\sigma) \tag{2.4}$$

where σ is the standard deviation, k is a factor relating the radii of the standard deviations of the two Gaussian functions G and τ changes the relative weighting between the larger and smaller Gaussians. As the resulting image has too much noise then, before noise removal, the $XDoG$ result $(Bxdog)$ is combined with the result from the Otsu binarization $(binOtsu)$ using Equation (2.5).

$$\text{If } abs(Bxdog - binOtsu) = 255 \text{ Then Bxdog } \leftarrow 0 \tag{2.5}$$

This combination is used to enhance the XDoG result without increasing the amount of noise. The noise from XDoG is removed using Otsu binarization result as a reference mask. The idea is to keep, in the cleaned image, only regions in XDoG image that satisfies at least one of two conditions: have more than 20 black pixels (ink) in size or match at least one black pixel in the Otsu binarized image

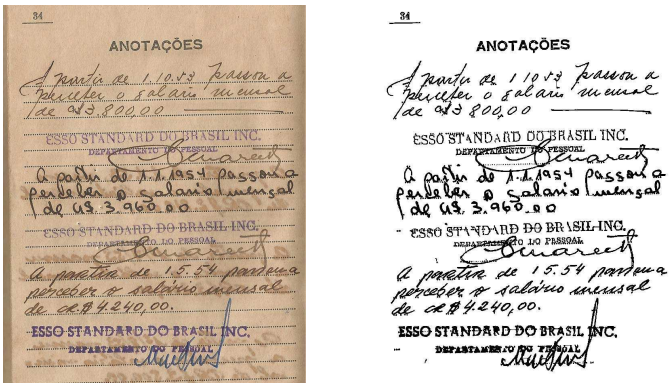


Figure 4. (left) Original image and (right) the result obtained by Roe and Mello algorithm.

3. Structural Contrast Based Thresholding Algorithms

The use of local contrast to improve the results of thresholding was used first with satisfactory results by Bernsen (1986). Other approaches came with the definition of a contrast image - a new way to represent an image with a better separability between text and background, making it easier to correctly classify the regions between both.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

In Su et al. (2010), it was proposed a thresholding algorithm for historical document images using contrast images to detect the border of the stroke. In this case, contrast is evaluated as the difference between the maximum and minimum intensity in a region. It was shown that the contrast evaluated through the absolute difference of the image, inside a local window, is sensible to contrast and brightness variations in an image. In order to compensate these variations, the authors proposed a normalization so that the contrast image is then evaluated as:

$$C(x,y) = \frac{I_{max}(x,y) - I_{min}(x,y)}{I_{max}(x,y) + I_{min}(x,y) + \varepsilon} \tag{2.6}$$

where $I_{max}(x,y)$ and $I_{min}(x,y)$ are the maximum and minimum local intensities in a window centered on the pixel (x,y). ε is just a small value used to avoid a division by zero. Otsu's thresholding algorithm was applied to the contrast image to detect high contrast pixels.

In the classification process, to each pixel of the original image, it is counted the number of text pixels, N_t, in a local window j in the bi-level image generated by Otsu. The pixel from the original image is considered a candidate to text pixel if N_t is greater than a predefined threshold (N_{min}). The classification is as follows:

$$R(x,y) = \begin{cases} 1, & \text{if } N_t \geq N_{min} \text{ and } I(x,y) \leq E_\mu + E_\sigma/2 \\ 0, & \text{otherwise} \end{cases} \tag{2.7}$$

where

$$E_\mu = \frac{\Sigma neighbors I(x,y) \times (1 - E(x,y))}{N_t} \tag{2.8}$$

and

$$E_\sigma = \sqrt{\frac{\Sigma neighbors((I(x,y) - E\mu) \times (1 - E(x,y)))^2}{2}} \tag{2.9}$$

$I(x,y)$ is the value of the pixel (x,y) of the original image. $E(x,y)$ is the value of the same position in the bi-level image created by Otsu. The method requires two parameters: the size of the window and the expected amount of ink pixels N_{min} inside the window. It is suggested to have a window with size greater than the stroke width.

Valizadeh and Kabir (2012) reported that the mapping of objects in an appropriate feature space can make a precise classification of pixels. Based on this, they have defined a mapping into a feature space that could separate text pixels from paper pixels. The proposal is divided into four steps: feature extraction, feature space partitioning, classification of the partitions, and pixel classification.

In the feature extraction phase, as the more relevant information is in the text, the most important features take into account structural features of the text, the structural contrast (SC) was created as (see Figure 5):

$$SC(x,y) = max_{k=0}^{3}\{min[M_{SW}(P_k), M_{SW}(P_{k+1}), M_{SW}(P_k')], M_{SW}(P_{k+1}')\} - I(x,y) \tag{2.10}$$

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

where

$$M_{SW}(P_k) = \left(\frac{\Sigma_{i=-SW}^{SW} \Sigma_{j=-SW}^{SW} I(P_{kx} - i, P_{ky} - j)}{(2 \times SW + 1)^2} \right) \tag{2.11}$$

with $I(x,y)$ being the intensity of pixel (x,y), P_{kx} and P_{ky} are P_k coordinates and SW is the stroke width as defined in Valizadeh et al. (2009). Thus, the pixel (x,y) is mapped into a 2D feature space, $A = [A1, A2]$, where $A1 = SC(x,y)$ and $A2 = I(x,y)$. The level of separability reached by the structural contrast can be analyzed in Figure 6-left.

P_1	P_2	P_3
$(P_4')P_0$	(x,y)	$P_4(P_0')$
$(P_3')P_7$	$P_6(P_2')$	$P_5(P_1')$

Figure 5. Neighbors pixels used to create the structural contrast.

In the space partitioning phase, a 2D histogram is evaluated from A. The mode association clustering algorithm (Li et al., 2007) is applied to the histogram. This technique partitions the feature space (Figure 6-left) into N small regions (as in Figure 6-right).

Niblack's local thresholding algorithm (Niblack, 1986) was proposed as the method to label the N regions. Suppose that $IM_{Niblack}$ is a bi-level image generated from the original image using Niblack's thresholding algorithm. To classify a region R_i, the total amount of pixels classified as text (N_t) or background (N_b) in the bi-level image are counted and the classification runs as follows:

$$R_i = \begin{cases} text, & \text{if } N_t(R_i) > N_b(R_i) \\ background, & \text{otherwise} \end{cases}$$

After this process, the feature space has just two regions as in Figure 8-right, defining the final bi-level image in a new thresholding operation based on this feature space. Figure 7 presents a sample image and the result after the application of Valizadeh and Kabir's algorithm.

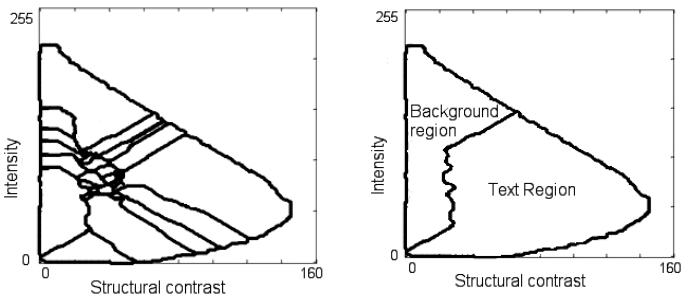


Figure 6. (left) Feature space partitioned into small regions and (right) these small regions are grouped as text or background edges.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 50 of 404

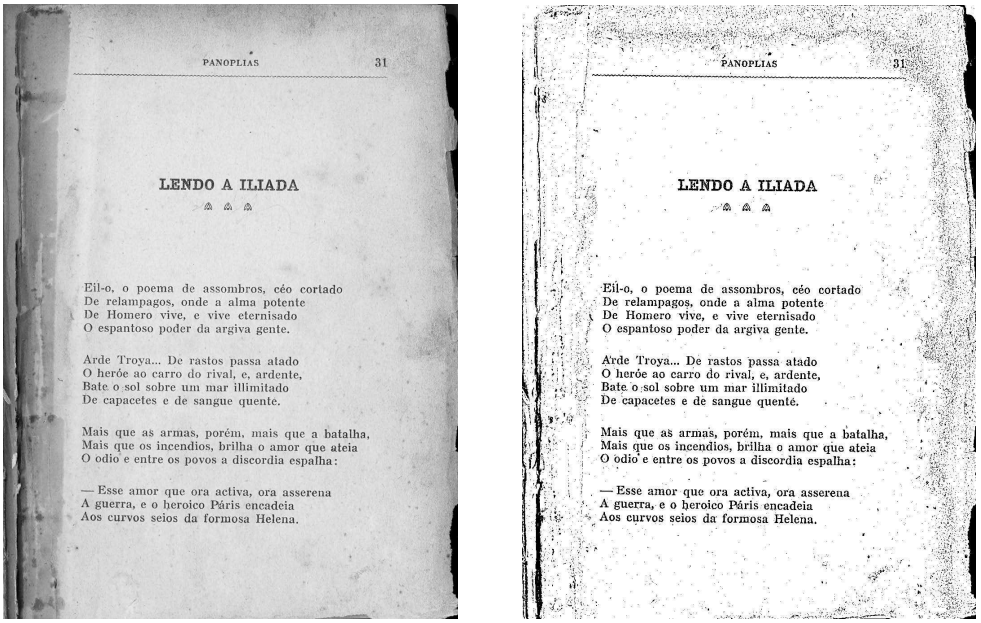


Figure 7. (left) Original old document and (right) resultant image by Valizadeh and Kabir.

It was noticed two situations where Valizadeh-Kabir's algorithm does not work properly:

- Sometimes, the structural contrast may not improve the separability between ink and paper. This usually happens when the ink has faded and its color becomes very close to the colors of the background; or, with the same consequences, there are smudges in the paper that darken it to turn its colors close to the ink.
- There is a high dependency on the results of Niblack's algorithm. This algorithm evaluates the average and standard deviation of the colors in a window and relates them through a variable k. The authors suggested $k = -0.2$ for all images. It is the same value for any image which is not reasonable and it is easy to find counter examples.

These issues led to the development of a new algorithm (Arruda-Mello) based on the work of Valizadeh and Kabir and published in Arruda and Mello (2014). The original method of Valizadeh and Kabir is used jointly with a new algorithm which creates a so-called "weak image" (an image where the major goal is background removal which can lead to the removal of some ink elements).This is done by the use of a normalized structural contrast image (SC_{Norm}):

$$SC_{Norm}(x,y) = \frac{M(x,y)_{max} - I(x,y)}{M(x,y)_{max} + M(x,y)_{min} + \varepsilon} \tag{2.12}$$

with

$$M(x,y)_{max} = max_{k=0}^{3}\{M(x,y)_{min}, M_{sw}(P'_{k-1})\} \tag{2.13}$$

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 51 of 404

and

$$M(x,y)_{min} = min[M_{SW}(P_k), M_{SW}(P_{k+1}), M_{SW}(P_k^{'})] \tag{2.14}$$

$I(x,y)$ is the gray value of the pixel $p(x,y)$ and is an infinitely small positive number used to avoid a division by zero. The neighborhood used for evaluation of the structural contrast is the same presented in Figure 7. $M_{SW}(P_k)$ is the average of the pixel intensities inside a window with center at (x,y), evaluated as previously defined in Equation (2.11).This normalization enhances the text regions and softens the effects of contrast/brightness variations between text and background.

The normalized structural contrast, however, does not have good results for regions with very low contrast. To improve it, SC_{Norm} image is combined with SC image to compensate this problem:

$$SC_{Comb}(x,y) = \alpha \times SC_{Norm}(x,y) + (1-\alpha) \times SC(x,y) \tag{2.15}$$

with $\alpha = (\sigma/128)^{\gamma}$, where σ is the standard deviation of the document image intensity and is a pre-defined parameter as proposed in Su et al. (2013). For a 256 gray level image, $\alpha \in [0,1]$ for any value of $\gamma > 0$. Both SC_{norm} and SC_{comb} are combined:

$$SC_{Mult}(x,y) = SC_{Norm}(x,y) \times SC_{Comb}(x,y) \tag{2.16}$$

Following there are two binarization processes both based on Valizadeh and Kabir method. As said before, one creates a "weak image", i.e., an image with possible loss of ink elements, while the other creates a "strong" image. The difference between them is that the first is created using SC image and the second is created using SC_{Mult}. They have different settings for the Niblack phase. A final step (the post-processing) is applied to the weak image restoring lost strokes based on the strong image. Figure 8 presents a sample image and the results generated by Valizadeh and Kabir and Arruda-Mello.

4. A Visual Perception Approach

A different approach for document image binarization is proposed in Mesquita et al. (2014). This approach is neither local nor global. In several cases, as it was shown in the previous sections, a thresholding algorithm reaches the separation of background and foreground through the enhancement of the foreground (the ink). This was clear in methods that are edge or local contrast based. Through a visual perception approach, it is possible to follow a different path.

For document images, the major objective of thresholding is the separation of ink and paper. It is also possible to reach this goal by enhancing the background. If we know what colors belong to the paper, we can also know what are the ones that belong to the ink. This is the main idea behind the method presented in Mesquita et al. (2014).

As we go far from an object, we lose the perception of details of that object just as corners become less sharp and more rounded. These effects are associated to distance

Cerence Operating Company, Ex. 2014
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 52 of 404

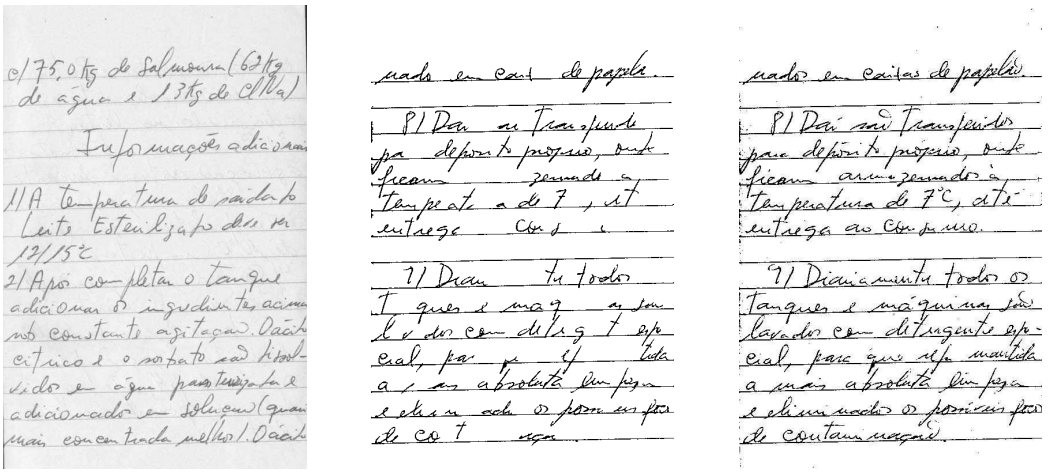


Figure 8. (left) Original image and its bi-level version created by (center) Valizadeh and Kabir and (right) Arruda-Mello.

perception (Goldstein and Brockmole, 2013). So if one goes far from a document, the details of the document will not be perceived anymore; in this case, the text or the ink part. However, the major colors that belong to the background will still be perceived. Figure 9 shows a simulation of what is expected to happen in this situation. It can be seen that as the observer moves away from the document, it fails to see the text. But the smudges of the paper are still visible.

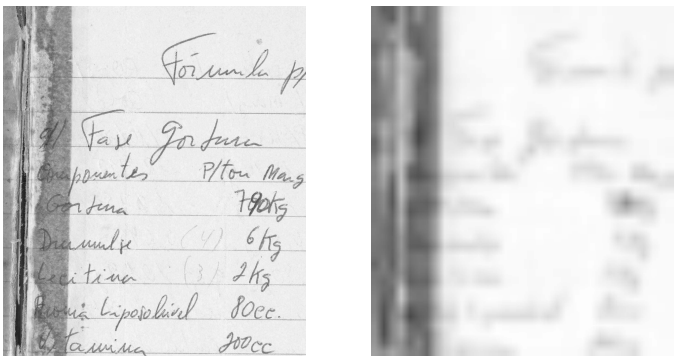


Figure 9. (left) Original image with smears; (right) a simulation of what is perceived at distance: although the text is not seen anymore, the marks of the smears are still perceived.

With this idea in mind, the method simulates the increasing of the distance between observer and document image through the use of resizing and morphological operations. Other operations are also applied as histogram equalization. As different stroke widths require different distances, the method starts by evaluating the thickness of the character so that the correct distance can be simulated. The stroke width is estimated as the median of all the nearest edge pixels found by the application of Sobel's operator (in the vertical direction) to the original image. Snellen's acuity test is the inspiration for the definition of the distance required to do not perceive the estimated ink anymore. Snellen visual acuity

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 53 of 404

test evaluates an individual's ability to detect a letter by measuring the minimum angle of resolution (the smallest target estimation in angular subtends that a person is capable of resolving).

In details, the method (called POD - Perception of Objects by Distance) works as follows:

1. Distance estimation based on stroke width;
2. Two morphological operations of closing are applied to the original image with disk as structuring elements (to achieve the rounded corners of objects);
3. Downsize the image to the size associated to the estimated distance (the size of the image that is formed on the observer's retina);
4. Resize back the previous image to the original size;
5. The absolute difference between the resized image and the original one is evaluated;
6. Dark pixels of the difference image are converted into white (as they represent perfect match of tones from the background);
7. All non-white pixels are assigned its complementary color;
8. Histogram equalization is applied.

These steps create a grayscale image that still needs to be binarized. However, although in grayscale, it is mostly composed just by background pixels; a fixed cut-off value, in general, already provides a good result. However, to guarantee a better quality image, a specific approach for binarization is also proposed. Otsu's thresholding algorithm (Otsu, 1979) and K-means clustering algorithm (MacQueen et al., 1967) are applied separately to the image generated after the 8th step presented before. A transition map is applied to the image produced by K-means in order to identify the text lines. These text lines are then used as a reference to clean the image produced by Otsu. A composition between this Otsu image and the K-means image creates the final bi-level image. Figure 10 illustrates the final result of the application of the algorithm on the image of Figure 9-left.

Most part of the algorithm is based on the application of image processing operations. However, one major step is focused in an aspect related to human vision. Thus, this step is being presented in more details. We are talking about the first step, the distance estimation based on stroke width. As explained before, this estimation is the core of the algorithm as the original idea comes exactly from what is perceived by the human visual system as the distance between observer and object increases.

The objective of this step is to lose the information about the ink so that just the pattern of the paper is perceived. It is natural to consider the stroke width as a feature to define the required distance. The stroke thickness on the image is estimated through the application of Sobel's edge detector in the vertical direction. It is measured, for each edge pixel, its distance to the nearest edge pixel in a horizontal direction. Most of the points detected by an edge detector (in the document image) may belong to the edge of a character; on the other hand, the edge detector usually detects some points that do not belong to the edge of

Document Ref. No. 2
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 54 of 404

a character, like edge points that belong to a smudge region, for example. The thickness of the characters is defined as the median of all the nearest edge pixels distances calculated. One weakness of the method is that just one stroke width is considered in an image; no variation is considered.

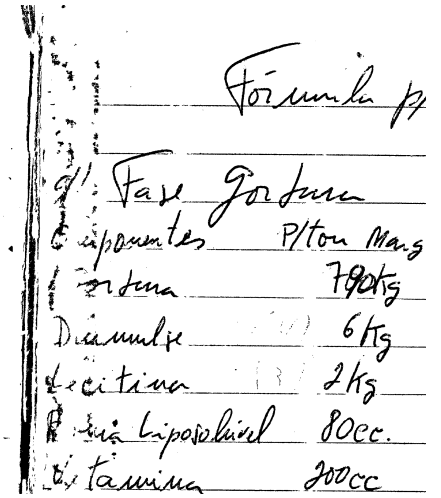


Figure 10. Final result of the application of perception based binarization algorithm on the image of Figure 9-left.

With the estimated stroke width, the distance can be evaluated. This is proposed with the inspiration of Snellen's acuity test. In this test, an observer is placed in front of a flowchart by certain distance. The flowchart has letters with different complexities and it estimates the individual's ability to recognize a letter by measuring the minimum angle of resolution (MAR): the smallest target estimate in angular subtends that the observer can perceive. Snellen acuity test is based on the standard that, to produce a minimal image in the retina, an object must subtend a visual angle of one minute of arc, 1'. As a consequence, as characters used in the test are 5 rows-high, each row subtends an angle of 1', the angle subtended by the characters is equal to 5'. Due to contrast variations between the real test and what it is displayed in a computer, it was considered a 3' angle. Thus, to define how far the image must be from the observer so that the ink is not perceived anymore, it is evaluated at which distance an object of the size of the estimated stroke thickness subtends an angle of 3'.

The needs for a perfect setting of the parameters of the algorithm motivated another study presented in Mesquita et al. (2015). In the case, the algorithm presented in Mesquita et al. (2014) is dependent on three parameters: the minimum angle resolution and the radiuses of both disks used in the closing morphological operation. For the radiuses, instead of considering them as parameters, it was used the difference between them. So, just two variables need to be optimized. I/F race algorithm (Birattari et al., 2010) is used to find the best solution to the problem. This algorithm with the best settings was submitted to H-DIBCO 2014 contest (Ntirogiannis et al., 2014) and ranked the first place.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

5. Graphical Documents

It is usual to think of documents as the standard “white” paper (in a sense that it is a paper with no previous element besides, maybe, guide lines) and the ink of the text (hand or type written). However, there are several different types of documents, considering as information stored in paper. For example, maps, floor plans, postcards, blueprints, are documents but with very different features if compared to a usual letter. Even a letter can be written on a letterhead which adds a graphical element to its contents. These types of documents are being called as Graphical Documents. They are documents where there is also some importance in the graphical information presented in it.

There are several applications to this kind of documents. Some of them are common to all these different types of graphical documents (as segmentation); others are more specific (as raster vector conversion for topographical maps).

A topographic map can be understood as a representation of a landscape which can contain contour lines, relief, constructions and natural features (as forests and lakes). Usually, maps contain its description in text regions. Other elements can also be found as illustrations or frames; these are very common in old maps. These features can also be found in floor plans, making these two types of documents very similar in some sense. When talking about old maps, blue prints or plans it is also usual to find them drawn in texturized or very soft papers (as butter paper) which make them more susceptible to degradation. One of the differences is that, in general, topographic maps are drawn by dark pens and sometimes painted with different colors to represent different kinds of regions; floor plans, however, can be drawn by pencil. Through time, the paper deteriorates (by the action of insects, fungi, humidity or just because of its natural fragility) and the ink can also fade away. Figure 11 presents (left) part of an old map and (right) part of an old floor plan. They are presented zooming in so that details can be better perceived.

From the images of Figure 11 it is possible to observe the following aspects:

1. The variations in the angles of the text;
2. The variations in font size and type;
3. The presence of overlapping (text over drawing);
4. The texturization of the floor plan paper.;

As an application of image processing of maps and plans we can fund the automatic indexation of them which can be reached with a more clean map. For this, the first step is binarization. This is not a simple task because of the degradation of the paper, folding marks (as the original maps and floor plans have in general high dimensions), damages, and so. For floor plans, some of them are drawn by pencil which let the strokes very clear. In Daniil et al. (2003), it is presented a study on scanning settings for digitization of old maps.

In Shaw and Bajcsy (2011), it is introduced a segmentation algorithm for automatic identification of regions in a map using reference regions in another map. The method makes a perfect match even if there is some level of differences. The authors also presented a map scale estimation method to evaluate the real area of the region according to the scale.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 56 of 404

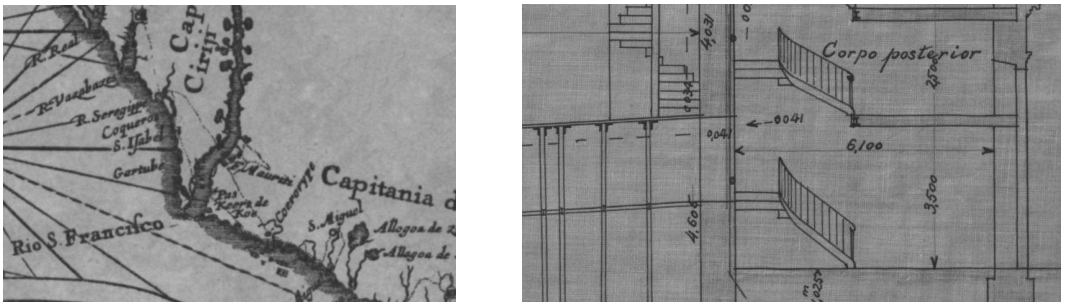


Figure 11. Zooming into (left) an old map (uncertain date) and (right) an old floor plan (dated from 1871).

Another automatic region matching applied to different maps is proposed in El-Hussainy et al. (2011).

Leyk and Boesch (2010) proposed a solution to color discrimination applied to low quality images of archival maps from the 19th century. The method proposed generates color layer prototypes by clustering; then it produces homogeneous color layer seeds. The connected regions are expanded based on region growing and the layers are segmented by filtering.

Filtering, clustering, statistical classification, edge detection and a local contrast enhancement are the major steps of the raster vector conversion method presented in Dezso et al. (2009).

A semiautomatic method for contour line extraction and 3D model construction is introduced in Ghircoias and Brad (2011). It is applied to topographic maps, requiring user intervention in two stages for adjustment of the settings. The user has to manually correct broken lines of the map. Segmentation is achieved by clustering, skeletonization and gap filling. Color quantization and noise removal are also applied. Vectorization and contour line interpolation (to create a 3D map of the terrain) are used to create the elevation model. The authors claim that such a system has to have a specialized user for parameter adjustment.

There are few works specific on floor plans when we talk about segmentation. One of them is presented in Ahmed et al. (2011). It works with typewritten text because the size of the text must be uniform inside a text region. This is not the common case in old document image processing. The method is not suitable for application to maps. It considers that the initial image was already binarized.

In Mello and Machado (2014), it is proposed a method for topological map and floor plan segmentation. The method is divided into two parts that can run in parallel; their results are combined to generate the final image. The final goal is the separation of text and drawings. As usual, the first step is the binarization of the original grayscale image. Figure 12 and Figure 13 illustrate how this step is important to the following parts of the algorithm. It presents the samples of Figure 11 and the results of different thresholding approaches (Valizadeh-Kabir, Arruda-Mello and POD).

In Figure 13, the images are presented larger so that the differences can be better perceived. The more evident differences can be seen in the superior right part of the figures.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

Valizadeh-Kabir algorithm is sensitive to the presence of the texturized paper so that some noise (remains of the texture) is present. The images produced by Arruda-Mello and POD are cleaner but POD's image has a better preservation of the stroke width as is observed in the comparison of the handwritten text "Corpo posterior" (in Portuguese).

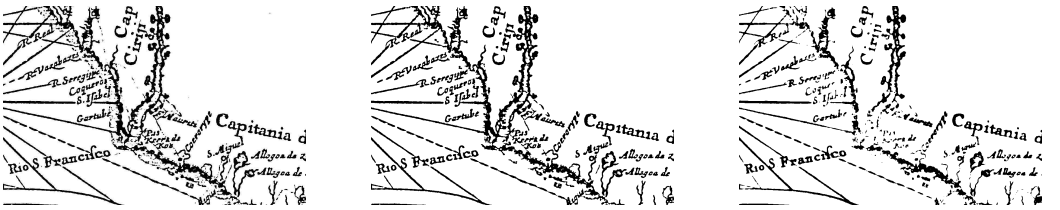


Figure 12. Sample map of Figure 11-left binarized by: (left) Valizadeh-Kabir, (center) Arruda-Mello and (right) POD.

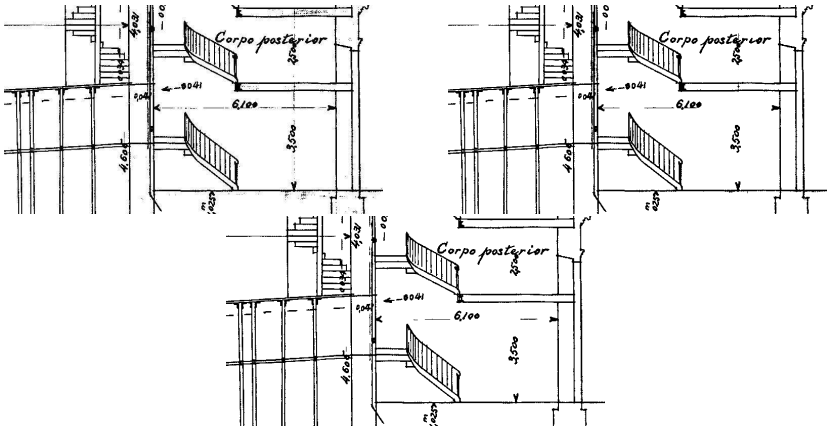


Figure 13. Sample floor plan of Figure 11-right binarized by: (top-left) Valizadeh-Kabir, (top-right) Arruda-Mello and (bottom) POD.

6. Thresholding Evaluation

One of the major problems of any new approach to solve a problem is how to show that this approach is better than what was already done. In some domains where the challenge is to develop faster algorithms, it is simple to measure a result. However, for thresholding this is still a problem. Even if you have a typewritten document, and the result of an optical character recognition tool could be used to measure the quality of your thresholding algorithm, there are issues that can be observed. For thresholding, any known evaluation strategy requires a gold standard (or ground truth). It is the expected best solution for your image. For document images, this could be the expected text file or the expected bi-level image. The major problem is how to create this gold standard and how to use it for comparison.

For a type written document, the ground truth can be a text file. The final bi-level image can be submitted to an optical character recognition tool and the resultant text file can be compared to the ground truth text file. In this case, someone has to have made the

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 58 of 404

transcription of the original document image into text. This is quite a problem when you have thousands of documents to transcribe as in an archive of old documents.

In the case of text analysis, text similarity algorithms are used for comparison. One of the most common metrics is the Levenshtein distance which measures the total amount of changes (insertions, deletions or substitutions) required to change a word into another. More robust approaches are presented in Gomaa and Fahmy (2013), a survey on text similarities methods. As our focus is image processing we are not going deeper into this line.

For binary document images, the problem is also complex and it also begins with the ground truth generation. Figure 14 makes the problem clear. With a grayscale image, there are two well defined regions: the inside of the character (the right white square in the ink) and the outside of the character (the left white square in the paper). However, there is a third area in which this classification is not so clear. It is the frontier between ink and paper; the region where the digitization process implies an aliasing between ink and paper areas in order to make the final image more pleasant to the human perception. And this is the area that can create difficulties in the ground truth production, possibly generating different responses.

One of the possible solutions is to use an edge detector algorithm (as Canny (Canny, 1986)) and let it detect the border of the characters. A sample result can be seen in Figure 15; it is possible to see that the result of the algorithm (the black edge) is not what we could call the best solution. However, due to the multiplicity of solutions, even a supervised approach would not reach a unique solution. More about the construction of ground truth images can be found in Ntirogiannis et al. (2008).

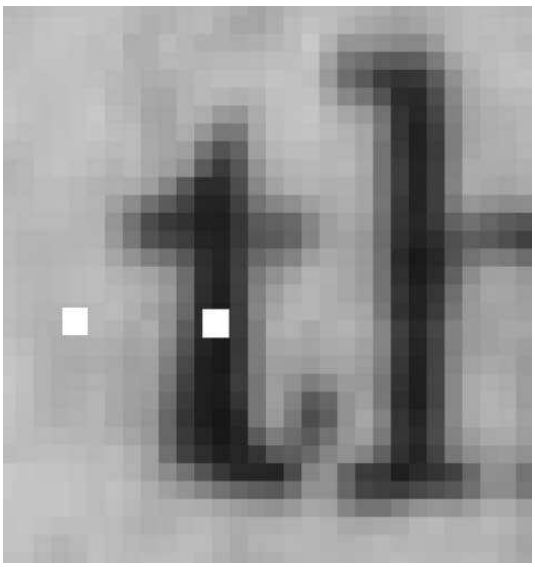


Figure 14. There is a fuzzy area between the certain paper and certain ink (left and right white squares respectively). It is not clear in this area which pixels belong to paper or ink.

With the ground truth images in hands, the next step is to determine the quality of a binarization algorithm and for this, a quantitative assessment is needed. The following measures, described in Ntirogiannis et al. (2014, 2008), can be used to get such quantitative

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 59 of 404

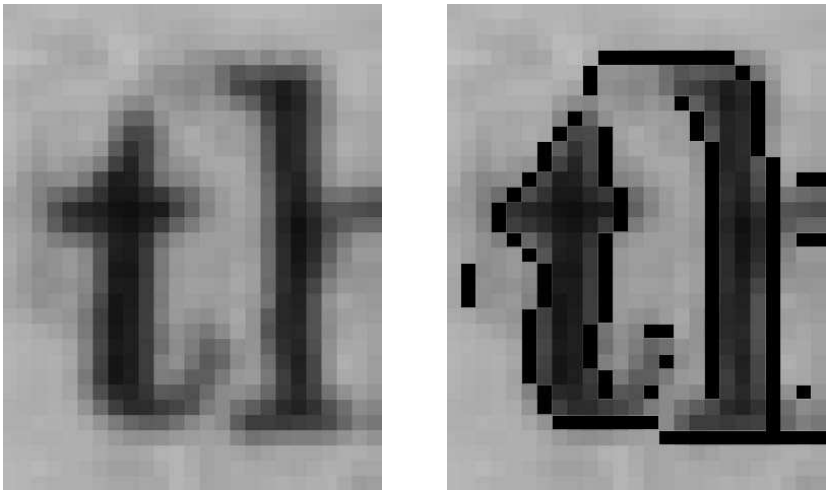


Figure 15. (left) Zooming into an old document and (b) the borders of the characters as detected by Canny's algorithm (in black).

estimates:

- Precision
- Recall
- Accuracy
- Specificity
- F-Measure
- Misclassification penalty metrics (MPM)
- Peak-signal to noise ratio (PSNR)
- Negative rate metric (NRM)

Before the description of the measures, some definitions necessary in the context of document imaging are presented:

- True positive (TP): the number of pixels correctly classified as ink;
- True negative (TN): number of pixels correctly classified as paper;
- False positive (FP): number of pixels that are part of the paper but are wrongly classified as ink;
- False negative (FN): number of ink elements classified as paper.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 60 of 404

For the use of these measures, it is necessary to have a ground truth reference image.

Precision. Is the fraction of retrieved instances that are relevant and is defined by Equation (2.17):

$$Precision = \frac{TP}{TP+FP} \tag{2.17}$$

A good algorithm must have $Precision \cong 1$ and for this is necessary that FP tend to zero meaning few errors.

Recall. Also known as sensitivity, is the fraction of true positives that are retrieved. Recall is defined by Equation (2.18):

$$Recall = \frac{TP}{TP+FN} \tag{2.18}$$

A good algorithm must have $Recall \cong 1$ then FN must tend toward zero.

Accuracy. Is the degree of approximation of a measured value of deemed correct, such as the ground truth, for example (Joseph et al., 2012). Accuracy is defined by Equation (2.19):

$$Accuracy = \frac{TP+TN}{P+N} \tag{2.19}$$

where:

$$P = TP+FN \text{ and } N = FP+TN \tag{2.20}$$

Specificity. Is also called the true negative rate and measures the proportion of negatives that are correctly identified as such. Specificity is defined by Equation (2.21).

$$Specificity = \frac{TN}{FP+TN} \tag{2.21}$$

A good algorithm must have $Specificity \cong 1$ e, then FP must tend toward zero.

F-Measure:. Is the Precision and Recall weighted harmonic mean, as defined by Equation (2.22).

$$FM = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{2.22}$$

Misclassification penalty metric (MPM). Evaluates the prediction against the ground and misclassified pixels are penalized by their distances from the ground truth object's border. The calculation of the MPM is given by Equation (2.23).

$$MPM = \frac{MP_{FN} + MP_{FP}}{2} \tag{2.23}$$

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 61 of 404

where:

$$MP_{FN} = \frac{\Sigma_{i=1}^{FN} d_{FN}^{i}}{D} \text{ and } MP_{FP} = \frac{\Sigma_{j=1}^{FP} d_{FP}^{j}}{D} \tag{2.24}$$

and d_{FN}^{i} e d_{FP}^{j} are the distances of the i-th false negative j-th false positive pixel from the ground truth contour. The normalization factor D is the sum over all pixel to contour distances of the ground truth. An algorithm with low MPM score means that it is good for object's boundary identification.

Peak Signal-to-Noise Ratio (PSNR). It is a measure of how an image is similar to others and, the larger the $PSNR$ value, greater is the similarity between them. Considering two images with dimensions $M \times N$, the $PSNR$ is defined by Equation (2.25):

$$PSNR = 10 \times log \frac{C^2}{MSE} \tag{2.25}$$

where MSE (Mean Square Error) is given by Equation (2.26):

$$MSE = \frac{\Sigma_{x=1}^{M} \Sigma_{y=1}^{N} (I(x,y) - I'(x,y))^2}{M \times N} \tag{2.26}$$

and C is the maximum color intensity (255 for a 8 bits grayscale image).

Negative Rate Metric (NRM). Is based on the discrepancy between pixels of the resulting image and the ground truth. The NRM combines the false negative rate (NR_{FN}) and the false positive rate (NR_{FP}) and is defined by Equation (2.27):

$$NRM = \frac{NR_{FN} + NR_{FP}}{2} \tag{2.27}$$

where:

$$NR_{FN} = \frac{N_{FN}}{N_{FN} + N_{TP}} \text{ and } NR_{FP} = \frac{N_{FP}}{N_{FP} + N_{TN}} \tag{2.28}$$

and N_{TP} represents the number of true positives, N_{FP} the number of false positives, N_{TN} the number of true negatives and N_{FN} the number of false negatives. Unlike $F - Measure$ and $PSNR$, the binarization quality is best for low NRM values. The ideal algorithm should have both FN and FP tending to 0 and Precision, Recall, Accuracy and Specificity tending to 1. This is a way to compare the results of thresholding algorithms as stated in Mello et al. (2008).

Conclusion

Document image thresholding, or binarization, is the initial step of most document image analysis system and refers to the conversion of a color or grayscale document image into

Cerence Operating Company, Ex. 2004
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 62 of 404

a bi-level image. The goal is to distinguish the text (ink) from the background (generally paper).

Although document image thresholding has been studied for many years, with several approaches proposed, it is still an unsolved problem. Different types of document degradation such as foxing, uneven illumination, image contrast variation, back-to-front ink interference etc. (as shown in Figure 1) are responsible for making this problem with a non-trivial solution.

In this chapter, some state-of-the-art algorithms, including the winner of the H-DIBCO 2014, were presented. The algorithms presented cover different approaches like edge based, structural contrast, algorithms to deal with graphical documents and a new approach based on the human visual perception system.

In Addition, it was discussed and presented measures for quantitative evaluations of the binarization algorithm's quality. For more information about the recent advances on thresholding and other document image processing techniques, we recommend to look for the proceedings of the following conferences: International Conference on Document Analysis and Recognition (ICDAR), International Conference on Document Engineering (DocEng), International Conference on Frontiers in Handwritten Recognition (ICFHR) and Workshop on Document Analysis Systems (DAS). We also recommend to follow the International Journal on Document Analysis Recognition (IJDAR). Look for the DIBCO and H-DIBCO contests annually in some of these previous conferences.

References

Ahmed, S., Weber, M., Liwicki, M., and Dengel, A. (2011). Text/graphics segmentation in architectural floor plans. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 734–738. IEEE.

Almazán, J., Gordo, A., Fornés, A., and Valveny, E. (2014). Segmentation-free word spotting with exemplar svms. *Pattern Recognition*, 47(12):3967–3978.

Arruda, A. and Mello, C. A. B. (2014). Binarization of degraded document images based on combination of contrast images. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 615–620. IEEE.

Ávila, B. T. and Lins, R. D. (2005). A fast orientation and skew detection algorithm for monochromatic document images. In *Proceedings of the 2005 ACM symposium on Document engineering*, pages 118–126. ACM.

Bernsen, J. (1986). Dynamic thresholding of grey-level images. In *International conference on pattern recognition*, volume 2, pages 1251–1255.

Birattari, M., Yuan, Z., Balaprakash, P., and Stützle, T. (2010). F-race and iterated f-race: An overview. In *Experimental methods for the analysis of optimization algorithms*, pages 311–336. Springer.

Brodić, D., Mello, C. A. B., Maluckov, Č. A., and Milivojevic, Z. N. (2014). An approach to skew detection of printed documents. *Journal of Universal Computer Science*, 20(4):488–506.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 63 of 404

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698.

Daniil, M., Tsioukas, V., Papadopoulos, K., and Livieratos, E. (2003). Scanning options and choices in digitizing historic maps. In *Proc. of CIPA 2003 International Symposium, Antalya, Turkey, September*.

Dezso, B., Elek, I., and Máriás, Z. (2009). Image processing methods in raster-vector conversion of topographic maps. In *Proceedings of the 2009 International Conference on Artificial Intelligence and Pattern Recognition*, pages 83–86.

El-Hussainy, M. S., Baraka, M. A., and El-Hallaq, M. A. (2011). A methodology for image matching of historical maps. *e-Perimetron*, 6(2):77–95.

Ghircoias, T. and Brad, R. (2011). Contour lines extraction and reconstruction from topographic maps. *Ubiquitous Computing and Communication Journal*, 6(2):681–691.

Goldstein, E. B. and Brockmole, J. (2013). *Sensation and perception*. Cengage Learning.

Gomaa, W. H. and Fahmy, A. A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13).

Gonzalez, R. C. and Woods, R. E. (2007). Image processing. *Digital image processing*, 2.

Joseph, A., Babu, J. S., Jayaraj, P., and KB, B. (2012). Objective quality measures in binarization. *International Journal of Computer Science and Information Technologies*, 3(4):4784–4788.

Leyk, S. and Boesch, R. (2010). Colors of the past: color image segmentation in historical topographic maps based on homogeneity. *GeoInformatica*, 14(1):1–21.

Li, J., Ray, S., and Lindsay, B. G. (2007). A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, 8(Aug):1687–1723.

Lu, S., Su, B., and Tan, C. L. (2010). Document image binarization using background estimation and stroke edges. *International Journal on Document Analysis and Recognition (IJDAR)*, 13(4):303–314.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Marr, D. and Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society of London B: Biological Sciences*, 207(1167):187–217.

Mello, C. A. B. (2012). *Digital document analysis and processing*. Nova Science Publishers, New York.

Mello, C. A. B., Sanchez, A., Oliveira, A., and Lopes, A. (2008). An efficient gray-level thresholding algorithm for historic document images. *Journal of Cultural Heritage*, 9(2):109–116.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 64 of 404

Mello, C. A. B. and Machado, S. (2014). Text segmentation in vintage floor plans and maps using visual perception. In *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*, pages 3476–3480. IEEE.

Mesquita, R. G., Mello, C. A. B., and Almeida, L. (2014). A new thresholding algorithm for document images based on the perception of objects by distance. *Integrated Computer-Aided Engineering*, 21(2):133–146.

Mesquita, R. G., Silva, R. M., Mello, C. A. B., and Miranda, P. B. (2015). Parameter tuning for document image binarization using a racing algorithm. *Expert Systems with Applications*, 42(5):2593–2603.

Niblack, W. (1986). An introduction to image processing.

Ntirogiannis, K., Gatos, B., and Pratikakis, I. (2008). An objective evaluation methodology for document image binarization techniques. In *Document Analysis Systems, 2008. DAS'08. The Eighth IAPR International Workshop on*, pages 217–224. IEEE.

Ntirogiannis, K., Gatos, B., and Pratikakis, I. (2014). Icfhr2014 competition on handwritten document image binarization (h-dibco 2014). In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 809–813. IEEE.

Otsu, N. (1979). A threshold selection method from gray-level histogram. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66.

Roe, E. and Mello, C. A. B. (2013). Binarization of color historical document images using local image equalization and xdog. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 205–209. IEEE.

Sánchez, A., Mello, C. A. B., Suárez, P. D., and Lopes, A. (2011). Automatic line and word segmentation applied to densely line-skewed historical handwritten document images. *Integrated Computer-Aided Engineering*, 18(2):125–142.

Sezgin, M. et al. (2004). Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic imaging*, 13(1):146–168.

Shaw, T. and Bajcsy, P. (2011). Automated image processing of historical maps. *SPIE Newsroom*.

Su, B., Lu, S., and Tan, C. L. (2010). Binarization of historical document images using the local maximum and minimum. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pages 159–166. ACM.

Su, B., Lu, S., and Tan, C. L. (2013). Robust document image binarization technique for degraded document images. *IEEE transactions on image processing*, 22(4):1408–1417.

Valizadeh, M. and Kabir, E. (2012). Binarization of degraded document image based on feature space partitioning and classification. *International Journal on Document Analysis and Recognition (IJDAR)*, 15(1):57–69.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 65 of 404

Valizadeh, M., Komeili, M., Armanfard, N., and Kabir, E. (2009). Degraded document image binarization based on combination of two complementary algorithms. In *Advances in Computational Tools for Engineering Applications, 2009. ACTEA'09. International Conference on*, pages 595–599. IEEE.

Winnemöller, H. (2011). Xdog: advanced image stylization with extended difference-of-gaussians. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Non-Photorealistic Animation and Rendering*, pages 147–156. ACM.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 66 of 404

In: Handwriting: Recognition, Development and Analysis ISBN: 978-1-53611-937-4
Editors: Byron L. D. Bezerra et al. © 2017 Nova Science Publishers, Inc.

Chapter 3

HISTORICAL DOCUMENT PROCESSING

***Basilis Gatos*[*], *Georgios Louloudis*[†], *Nikolaos Stamatopoulos*[‡] *and Giorgos Sfikas*[§]**
Computational Intelligence Laboratory,
Institute of Informatics and Telecommunications,
National Center for Scientific Research "Demokritos"
Athens, Greece

1. Introduction

Historical manuscript collections can be considered as an important source of original information in order to provide access to historical data and develop cultural documentation over the years. This chapter reports on recent advances and ongoing developments for historical handwritten document processing. It outlines the main challenges involved, the different tasks that have to be implemented as well as practices and technologies that currently exist in the literature. The focus is given on the most promising techniques as well as on existing datasets and competitions that can be proved useful to historical handwritten document processing research.

The main tasks that have to be implemented in the historical document image recognition pipeline, include preprocessing for image enhancement and binarization, segmentation for the detection of main page elements, text lines and words and, finally, recognition. In cases where optical recognition is expected to give poor results, keyword spotting has been proposed to substitute full-text recognition.

The organization of this chapter is as follows. Section "Preprocessing" gives an overview of document image enhancement and binarization methods while section "Segmentation" presents layout analysis, text line and word segmentation state-of-the-art techniques for historical handwritten documents. In section "Handwritten Text Recognition (HTR)" the focus is on the pure recognition task which can be accomplished on text line,

[*]E-mail address: bgat@iit.demokritos.gr.
[†]E-mail address: louloud@iit.demokritos.gr.
[‡]E-mail address: nstam@iit.demokritos.gr.
[§]E-mail address: sfikas@iit.demokritos.gr.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

word or character level. Finally, in section “Keyword spotting” recent advances on searching for a keyword directly on the historical document images are presented.

2. Preprocessing

The conservation and readability of historical manuscripts are often compromised by several types of degradations which not only reduce the legibility of the historical documents but also affect the performance of subsequent processing such as document layout analysis (DLA) and handwritten text recognition (HTR); therefore a preprocessing procedure becomes essential. Once an efficient preprocessing stage has been applied, the performance of processing systems is improved while at the same time preprocessed and enhanced documents become more attractive to users such as humanities scholars.

The term degradation has been defined by Baird (2000) as follows: “By degradation (or defects), we mean every sort of less-than-ideal properties of real document images”. On the basis of origin, degradations can be classified into three different categories. Historical document images may contain degradations due to (i) the image acquisition process as well as (ii) the environmental conditions and ageing (i.e. humidity, manipulation, unfitted storage). Specifically, concerning handwritten documents, (iii) the use of quill pens is also responsible for several degradations (i.e. seeping of ink from the reverse side, different amount of ink and pressure by the writer). According to this categorization, degradations of historical manuscripts can be:

(i) Speckles, salt and pepper noise, blurring, shadows, low-resolution artifacts, curvature of the document

(ii) Non-uniform background and illumination changes due to paper deterioration and discoloration, spots and poor contrast due to humidity, smudges, holes, folding marks

(iii) Faint characters, bleed-through, presence of large stains and smears

Taking into account the type of the enhancement methodology which should be applied, historical document image degradations are also categorized into background degradations, foreground degradations as well as global degradations (Drira, 2006). Concerning the first category, degradations consist of artifacts in the background (e.g. bleed-through) in which classification methods should be applied in order to separate these artifacts from the useful textual information. Foreground degradations affect textual information (e.g. faint characters) and should be restored by the enhancement procedure. Finally, the last category refers to degradations which affect the entire document, such as geometrical distortions, in which the enhancement stage is oriented towards modelling the image degradations. Examples of degraded historical manuscripts are depicted in Figure 1.

Several historical handwritten document image preprocessing techniques have been reported in the literature. Each of these techniques depends on a certain context of use and is intended to process a precise type of degradations or a combination of them. These techniques fall broadly into two main categories according to the type of the produced document image: (i) document image enhancement methods and (ii) document image binarization methods.

Document image enhancement methods aim to improve the quality of the original color or grayscale image. The produced document image after the enhancement procedure is also a color or grayscale image. On the other hand, document image binarization refers to the

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 68 of 404



Figure 1. Examples of degraded historical handwritten document images.

conversion of a color/grayscale image into a binary image. The main goal is not only to enhance the readability of the image but also to separate the useful textual content from the background by categorizing all the pixels as text or non-text without missing any useful information. Techniques of the former category are used also as a preparation stage for the binarization methods.

In the remainder of this section, the major enhancement and binarization techniques for historical handwritten documents will be presented along with the corresponding evaluation protocols.

2.1. Enhancement Techniques

As already mentioned, historical manuscripts suffer from several types of degradations. One of the most common degradations is the bleed-through effect and for this reason several enhancement techniques which focus on this type of effect have been reported in the literature. Bleed-through is caused by the seeping of ink from the reverse side, or it appears when the paper is not completely opaque (show-through). Consequently, text information from the back interferes with the text in the front page and the use of binarization techniques is often not effective since the intensities of the reverse side can be very close to those of

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 69 of 404

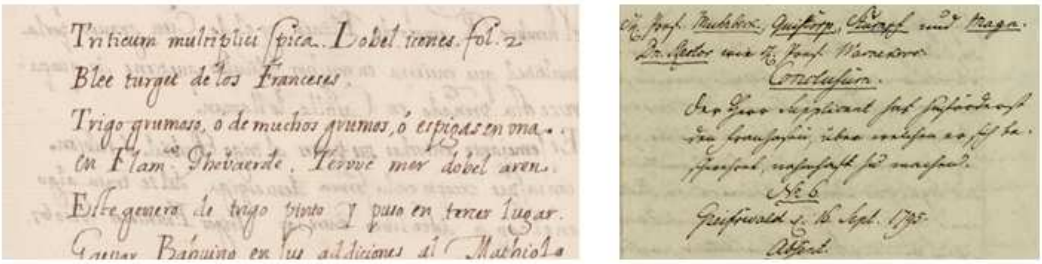


Figure 2. Examples of bleed-through degraded historical manuscripts.

the foreground text (see Figure 2).

The enhancement techniques which cope with the bleed-through effect can be divided into two categories according to the presence (or not) of the verso document image: (i) non-blind techniques in which both sides of the document image are available and (ii) blind techniques which process a single-side document image.

Non-blind techniques are mainly based on the comparison between the recto and verso pages in which a preliminary registration of the two sides is required. Tan et al. (2002) proposed a wavelet reconstruction process for iteratively enhancing the foreground strokes and smearing the interfering strokes. An improved Canny edge detector to suppress unwanted interfering strokes has been also used. However, the alignment of both images was done manually. In Tonazzini et al. (2007), the authors presented a non-blind method applicable to grayscale document images using a linear model based on the blind source separation (BSS) technique. Independent Component Analysis (ICA) and Principal Component Analysis (PCA) were employed in order to separate recto from verso information. This method requires a single, very fast, processing step, with no need for segmentation or inpainting. However, linear models despite their lower computation cost are not very suitable for the analysis of nonlinear problems. Another technique (Moghaddam and Cheriet, 2010) removes bleed-through effect using a variational approach. The variational model is adapted using an estimated background according to the availability of the verso side of the document image since it can be applied also as a blind technique. An advanced model based on a global control, the flow field, is introduced which helps to preserve the very weak edges, while at the same time achieving a high degree of smoothing and enhancement. The proposed model is robust with respect to noise and complex background.

In the case where the reverse side of the document image is not available, blind techniques are required in which only one document image is processed. Tonazzini et al. (2004) proposed a method which is based on the BSS technique and it takes advantage of the color image. The image is modeled as a linear combination of the interfering texts which are separated by processing multiple views of the image. If the color version of the image is available, three different views can be obtained from the red, green and blue image channels. In Drira (2006), a recursive non-supervised segmentation approach has been proposed which is based on the k-means algorithm. The dimension of the image is reduced and its data is decorrelated using PCA computed on the RGB color space. The stopping criterion for the proposed recursive approach has been determined empirically and set to a

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

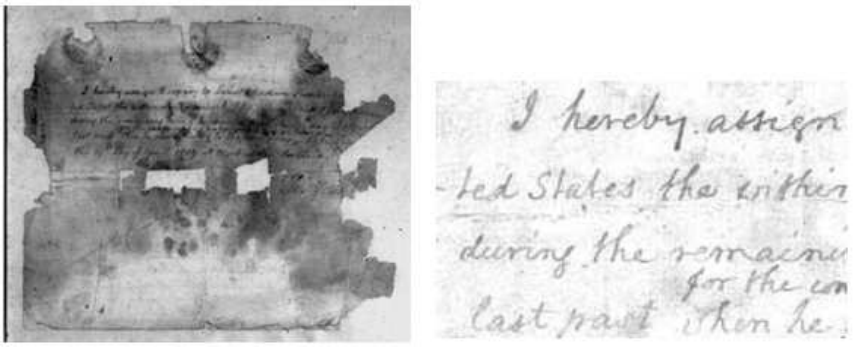


Figure 3. Example of an enhanced manuscript produced by Shi and Govindaraju (2004a) (a) Original image and (b) a portion of the enhanced image.

fixed number of iterations. Another blind approach has been proposed by Wolf (2010). It is based on separate Markov Random Field (MRF) regularization for the recto and verso side, where separate priors are derived from the full graph. The segmentation algorithm is based on Bayesian Maximum a Posteriori (MAP) estimation. Finally, Villegas and Toselli (2014) presented an enhancement method based on learning a discriminative color channel by considering a set of labeled local image patches. The user should point out explicitly for some sample pages which parts are bleed-through as well as which parts are clean text, with the aim that the method will be adapted to the characteristics of each document. The technique is intended to be part of an interactive transcription system in which the objective is obtaining high quality transcriptions with the least human effort.

All the above mentioned techniques focus on the correction of the bleed-through effect. Several other degradations have been addressed by enhancement methods. For example, Shi and Govindaraju (2004a) proposed a background light intensity normalization algorithm suitable for historical manuscripts with uneven background. A linear model is used adaptively to approximate the paper background. Then the document image is transformed according to the approximation to a normalized image that shows the foreground text on a relatively even background. The method works for grayscale as well as color images. An example of an enhanced manuscript produced by this method is depicted in Figure 3. In Gangamma et al. (2012), a restoration method was proposed in order to eliminate noise, uneven background and enhance the contrast of the manuscripts. The proposed method combines two image processing techniques, a spatial filtering technique and grayscale mathematical morphology operations. Furthermore, Saleem et al. (2014) proposed a restoration method in order to reduce the background noise and enhance the text information. A sliding window is applied in order to calculate the local minimum and maximum pixel intensities which are used for image normalization.

Finally, enhancement techniques which are based on the hyperspectral imaging system (HSI) using special equipment have been reported in the literature. HSI is useful for many tasks related to document conservation and management as it provides detailed quantitative measurements of the spectral reflectance of the document that is not limited to the visible spectrum. Joo Kim et al. (2011) proposed an enhancement strategy for historical documents captured by a hyperspectral imaging system. This method tries to combine an original RGB

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 71 of 404

image with images taken in Near IR range in order to preserve the texture of the image. Therefore an enhancement step is performed in the gradient domain which is dedicated to the removal of artifacts. In a similar way, Hollaus et al. (2014) presented an enhancement method for multispectral images of historical manuscripts. The proposed method is based on the Linear Discriminant Analysis (LDA) technique. LDA is a supervised technique and hence a labeling of training data is required. For this purpose, two different labeling strategies are proposed, which are both based on spatial information. One method is concerned with the enhancement of non-degraded image regions and the other technique is applied on degraded image portions. The resulting images are afterwards merged into the final enhancement result.

Although various enhancement techniques have been proposed, no standard performance evaluation methodology exists. Most of the evaluations concentrate in visual inspection of the resulted document image. The performance of these techniques is based on subjective human evaluation; hence objective evaluations among the different techniques cannot be obtained. For example, in Tan et al. (2002) the enhanced manuscripts were visually inspected in order to count the number of words that are fully restored. The performance of the system is measured in terms of precision and recall according to the total number of words in the original image.

Another strategy to evaluate enhancement techniques is the use of OCR as a means for indirect evaluation by comparing the OCR performance on original and enhanced images. However, in many cases, such as in historical handwritten documents, a meaningful OCR is not always feasible. In Tonazzini et al. (2007) and Wolf (2010), the authors presented restoration examples of historical manuscripts but they carried out the OCR evaluation on historical printed documents. On the other hand, in Villegas and Toselli (2014), Saleem et al. (2014) and Hollaus et al. (2014) the restoration performance is evaluated by means of HTR using historical handwritten datasets.

2.2. Binarization Techniques

Document image binarization techniques are usually classified in two main categories, namely global and local thresholding. Global thresholding methods use a single threshold value for the entire image, while local thresholding methods detect a local (adaptive) threshold value for each pixel. Global techniques are capable of extracting the document text efficiently in the case that there is a good separation between the foreground and the background. However, they cannot effectively handle historical handwritten document images with degradations, such as non-uniform background and faint characters.

Several historical binarization methods have incorporated background subtraction in order to cope with several degradations (see Figure 4). Gatos et al. (2006) proposed a method which estimates the background by taking into account the result of the adaptive Sauvola binarization method (Sauvola and Pietikinen, 2000) applied after a preprocessing step. The final threshold is based on the difference between the estimated background and the preprocessed image. Finally, a post-processing enhancement step is applied in order to improve the quality of text regions and preserve stroke connectivity. In a similar approach, in Lu et al. (2010), the background is estimated using an iterative polynomial smoothing procedure. Different types of document degradations are then compensated by

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 72 of 404

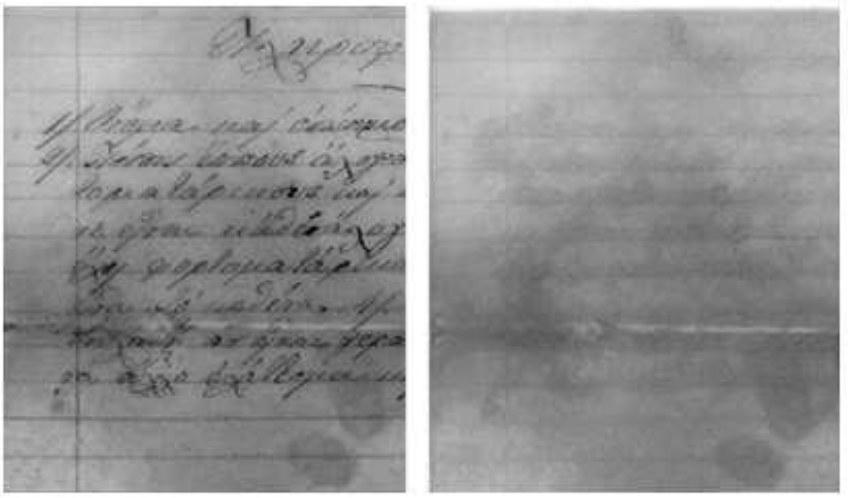


Figure 4. Background surface estimation using (Gatos et al., 2006) method (a) Original image and (b) background surface.

using the estimated document background surface. The original image was normalized and the text stroke edges were detected. Finally, the local threshold was based on the local number of the detected text stroke edges and their mean intensity. This method is based on the local contrast for the final thresholding and hence some bleed-though or noisy background components of high contrast remain. Another binarization method which is based on background subtraction is presented by Ntirogiannis et al. (2014a). The proposed binarization method was developed specifically for historical handwritten document images and it comprises several steps. In the first step, background estimation is performed using an inpainting procedure initialized by a local binarization method. In the sequel, image normalization is applied to correct large background variations. Then, a global and a local binarization method are applied to the normalized image and their results are combined at connected component level. Intermediate processing to remove small noisy connected components is also applied. This method could miss textual information in an attempt to clear the background from noisy components or bleed-through.

Edge-based techniques are another category of binarization methods which usually use a measure of the intensity changes across an edge (local contrast computation). For example, in Su et al. (2010) the image contrast is calculated (based on the local maximum and minimum intensity) and the edges are detected using a global binarization method. Compared with the image gradient, the image contrast evaluated by the local maximum and minimum has a nice property which makes it more tolerant to the uneven illumination and other types of document degradation such as smear. The document text is then segmented by using local thresholds that are estimated from the detected high contrast pixels within a local neighborhood window. This method is capable of removing the majority of the background noise and bleed-through but it is not so efficient in faint characters detection. An extension of this work is presented in Su et al. (2013) by the authors which addresses the over-normalization problem of the previous work. The proposed method is simple, robust and capable of handling different types of historical manuscript degradations with mini-

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 73 of 404

mum parameter tuning. It makes use of the adaptive image contrast that combines the local image contrast and the local image gradient adaptively and therefore is tolerant to the text and background variation.

A decomposition method has been presented by Chen and Leedham (2005) for thresholding degraded historical documents. Chen and Leedham proposed an algorithm which recursively decomposes a document image into subregions until appropriately weighted values can be used to select a suitable single-stage thresholding method for each region. The decomposition algorithm uses local feature vectors to analyze and find the best approach to threshold a local area. A new mean-gradient-based method to select the threshold for each subregion is also proposed. Moreover, multi-scale approaches have been used in some works in order to separate the text from the background. A grid-based modeling has been introduced by Farrahi Moghaddam and Cheriet (2010). This method is able to improve the binarization results and restore weak connections and strokes, especially in the case of degraded historical documents. Using the fast, grid-based versions of adaptive methods, multi-scale methods are created which are able to recover the text on several scales and restore document images with complex backgrounds that suffer from intensity degradation. The authors presented also an adaptive modification of the global Otsu binarization method, called AdOtsu. Finally, in the recent work (Afzal et al., 2015) the binarization problem is treated as a sequence learning problem. The document image is considered as a 2D sequence of pixels and in accordance to this, a 2D Long Short-Term Memory (LSTM) is employed for the classification of each pixel as text or background. The proposed method processes the information using local context and then propagates the information globally in order to achieve better visual coherence. It requires no parameter tuning and works well without any feature extraction. While learning methods require a large amount of training data and also similar type of images, this method works efficiently with limited amount of data.

Performance evaluation strategies of document image binarization techniques can be classified in three main categories: (i) visual inspection of the final result (Gatos et al., 2006), (ii) indirect evaluation by taking into account the OCR performance of the binary image with respect to character and word accuracy (Farrahi Moghaddam and Cheriet, 2010) and (iii) direct evaluation by taking into account the pixel-to-pixel correspondence between the ground truth and the binary image. Direct evaluation is based either on synthetic or real images. A performance evaluation methodology which focuses on historical documents containing complex degradations has been proposed by Ntirogiannis et al. (2013). It is a pixel-based evaluation methodology which introduces two new measures, namely pseudo-Recall and pseudo-Precision. It makes use of the distance from the contour of the ground truth to minimize the penalization around the character borders, as well as the local stroke width of the ground truth components to provide improved document-oriented evaluation results. In addition, useful error measures, such as broken and missed text, character enlargement and merging, background noise and false alarms, were defined that make more evident the weakness of each binarization technique being evaluated.

Furthermore, a series of document image binarization contests (DIBCO and H-DIBCO) have been organized in the context of the ICDAR and ICFHR conferences in order to identify current advances in document image binarization using established evaluation performance measures. DIBCO contests (Gatos et al. (2009), Pratikakis et al. (2011), Pratikakis

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 74 of 404

et al. (2013)) consist of handwritten and machine-printed document images and, on the other hand H-DIBCO contests (Pratikakis et al. (2010), Pratikakis et al. (2012), Ntirogiannis et al. (2014b)) contain only handwritten document images. The ground-truth binary images were created following a semi-automatic procedure. Tables 1- 3 illustrate performance evaluation results of several binarization methods which were mentioned above, using DIBCO 2009 (Gatos et al., 2009) and H-DIBCO 2010 (Pratikakis et al., 2010) datasets in terms of F-Measure, PSNR, Negative Rate Metric (NRM) and Misclassification Penalty Metric (MPM). The final ranking was calculated after sorting the accumulated ranking value for all measures. Concerning DIBCO 2009 dataset, which consists of handwritten and machine-printed document images, evaluation results using only the handwritten images are also presented. As the evaluation results indicate, the method developed by Ntirogiannis et al. (2014a) outperforms all the other techniques concerning the handwritten document images.

Table 1. Evaluation results using DIBCO2009 dataset.

Rank	Method	F-Measure (%)	PSNR	NRM $(x10^{-2})$	MPM $(x10^{-3})$
1	Su et al. (2013)	93.50	19.65	3.74	0.43
2	Lu et al. (2010)	91.24	18.66	4.31	0.55
3	Su et al. (2010)	91.06	18.50	7.00	0.30
4	Gatos et al. (2006)	85.25	16.50	10.00	0.70

Table 2. Evaluation results using only the handwritten images of DIBCO2009 dataset.

Rank	Method	F-Measure (%)	PSNR	NRM $(x10^{-2})$	MPM $(x10^{-3})$
1	Ntirogiannis et al. (2014a)	92.64	21.28	2.84	0.48
2	Su et al. (2010)	89.93	19.94	6.69	0.30
3	Lu et al. (2010)	88.65	19.42	5.11	0.34

Table 3. Evaluation results using H-DIBCO2010 dataset.

Rank	Method	F-Measure (%)	PSNR	NRM $(x10^{-2})$	MPM $(x10^{-3})$
1	Ntirogiannis et al. (2014a)	94.34	21.60	3.04	0.32
2	Su et al. (2013)	92.03	20.12	6.14	0.25
3	Lu et al. (2010)	86.41	18.14	9.06	1.11
4	Su et al. (2010)	85.49	17.83	11.46	0.37
5	Gatos et al. (2006)	71.99	15.12	21.89	0.41

3. Segmentation

Document segmentation is introduced in the first steps of the document processing pipeline and corresponds to the correct localization of the main page elements of a document. This step is further analyzed to the layout analysis, text line segmentation and word segmentation stages. All the abovementioned stages are very important since their success plays a

Cerence Operating Company, Ex. 2014
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 75 of 404

significant role to the accuracy of the final recognition result. This section is dedicated to the analytical presentation of the three stages, with respect to the challenges appearing on historical handwritten documents, the latest achievements found in the literature as well as evaluation results which reflect the level of maturity of each stage.

3.1. Layout Analysis

Layout analysis refers to the process of identifying as well as categorizing the regions of interest (e.g. text blocks, ruler lines, marginalia, figures, tables, drawings, ornamental characters) which exist on a handwritten document image. A reading system requires the detection of main page elements as well as the discrimination of text zones from non-textual ones in order to facilitate the recognition procedure. Historical handwritten documents do not have strict layout rules and thus, a layout analysis method needs to be invariant to layout inconsistencies, irregularities in script and writing style, skew, fluctuating text lines, and variable shapes of decorative entities (see Figure 5).



(a) (b) (c)

Figure 5. (a) Latin document of two columns with ornamental characters for each paragraph (Baechler and Ingold, 2011), (b) Arabic document with complex layout due to the existence of side-note text (Bukhari et al., 2012), (c) Latin document image with complex layout from the Bentham dataset (Gatos et al., 2014). Notice the existence of ruler lines, the stamp and page number on the top right as well as the deleted text on the first text line.

Layout analysis methods reported in the literature can be classified into two distinct categories, namely **bottom-up** and **top-down** approaches. Bottom-up methods start from small entities of the document image (e.g. pixels, connected components). These entities are grouped into larger homogeneous areas leading to the creation of the final regions of interest. On the contrary, top-down methods start from the document image and repeatedly split it to smaller areas according to specific rules which, finally, correspond to distinct regions of interest. An alternative taxonomy can be defined in the case that training data

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 76 of 404

exist. According to this taxonomy, there exists the category of **supervised methods** which assume the existence of an already annotated dataset serving as the training part used to train an algorithm for distinguishing the regions of interest. Methods that do not make use of any prior knowledge and thus no training is involved, are said to belong to the category of **unsupervised methods**.

Several layout methods for historical handwritten documents have been reported in the literature. Nicolas et al. (2006) proposed to use Markov Random Fields for the task of complex handwritten document segmentation and presented an application of the method on Flaubert's manuscripts. The authors report 90,3% in terms of global labeling rate (GLR) and 88,2% in terms of normalized labeling rate (NLR) using the Highest Confidence First (HCF) image labeling method on a set of 23 document images of the Flaubert's manuscripts. The task considered consists of labeling the main region of a manuscript i.e. text body, margins, header, footer, page number and marginal annotations. Bulacu et al. (2007) presented a layout analysis method applied on the archive of the cabinet of the Dutch Queen which consists of the generation of a course layout of the document by finding the page borders, the rule lines of the index table and the handwritten text lines grouped into decision paragraphs. Visual evaluation is performed due to lack of ground truth information on a dataset of 1040 document images showing encouraging results. Baechler and Ingold (2011) described a generic layout analysis system for historical documents. Their implementation used a so called Dynamic Multi-Layer perceptron (DMLP), which is a natural extension of MLP classifiers. The system was evaluated on medieval documents for which a multi-layer model was used to discriminate among 10 classes organized hierarchically.

Bukhari et al. (2012) introduced an approach which segments text appearing in page margins (see Figure 5b). A MLP classifier was used to classify connected components to the relevant class of text together with a voting scheme in order to refine the resulting segmentation and produce the final classification. The authors report a segmentation accuracy of 95% on a dataset of 38 Arabic historical document images. Asi et al. (2014), worked on the same Arabic dataset proposing a learning-free approach to detect the main text area in ancient manuscripts. They refine an initial segmentation using a texture-based filter by formulating the problem as an energy minimization task and achieving the minimum using graph cuts. This method is shown to outperform (Bukhari et al., 2012) achieving an accuracy of 98.5%.

Cohen et al. (2013) presented a method to segment historical document images into regions of different content. A first segmentation is achieved using a binarized version of the document, leading to a separation of text elements from non-text elements. A refinement of the segmentation of the non-text regions into drawings, background and noise is achieved by exploiting spatial and color features to guarantee coherent regions. The authors report approximately 92% and 90% segmentation accuracy of drawings and text elements, respectively, on a historical dataset of 252 pages. Gatos et al. (2014) proposed a text zone detection aiming to handle several challenging cases such as horizontal and vertical rule lines overlapping with the text as well as two column documents. The authors reported an accuracy of 84.7% for main zone detection on a dataset consisting of 300 pages.

A general remark concerning the abovementioned methods is that a direct comparison cannot be made in order to clearly understand which method is superior with respect

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 77 of 404

to the others. The main reason is that each work uses different data for evaluation, different evaluation metrics and, most importantly, different page elements are detected per method. Table 4 presents the categorization of the abovementioned methods to the different taxonomies described as well as the number of different page elements detected by each method.

Table 4. Categorization of state-of-the-art layout analysis methods

Method	Bottom-Up	Top-Down	Supervised	Unsupervised	No. Page Elements
Nicolas et al. (2006)	x		x		6
Bulacu et al. (2007)		x		x	5
Baechler and Ingold (2011)	x		x		6
Bukhari et al. (2012)	x		x		2
Asi et al. (2014)	x		x		2
Cohen et al. (2013)	x		x		2
Bukhari et al. (2012)	x		x		2
Gatos et al. (2014)		x		x	1

3.2. Text Line Segmentation

Text line segmentation which is the process of defining the region of every text line on a document image constitutes one of the most important stages of the handwritten text recognition pipeline. Results of poor quality produced by this stage seriously affect the accuracy of the handwritten text recognition procedure. Several challenges exist on historical documents which should be addressed by a text line segmentation method. These challenges include a) the difference in the skew angle between lines on the page or even along the same text line, b) overlapping and touching text lines, c) additions above the text line and d) deleted text. Figure6 presents one example on each of these challenges.

A very interesting survey covering the challenges, the categorization of existing methods as well as several open issues concerning the task of text line segmentation for historical documents has been introduced by Likforman-Sulem et al. (2007). In this survey, text line segmentation methods are said to fall broadly into four categories: i) Projection-based methods, ii) Smearing methods, iii) Grouping methods and iv) Hough transform based methods. A similar taxonomy can be found in the work of Louloudis et al. (2009). Recently, a fifth category of text line methods has arised since many researchers were motivated by the work of Avidan and Shamir (2007) which introduced the use of seams for treating the problem of image resizing. The main idea of the text line segmentation methods belonging to the seam based category concerns the use of an energy map which is used to determine seams that pass across and between text lines.

Projection-based methods include the work of Bar-Yosef et al. (2009). The method consists of two steps. The first step concerns the calculation of the local projection profile for each vertical stripe of the document image. The second step corresponds to the detection of local minima for each projection profile. The authors conducted experiments on 30 degraded historical documents. Evaluation was based on visual inspection for which it is reported a correct segmentation of 98%.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

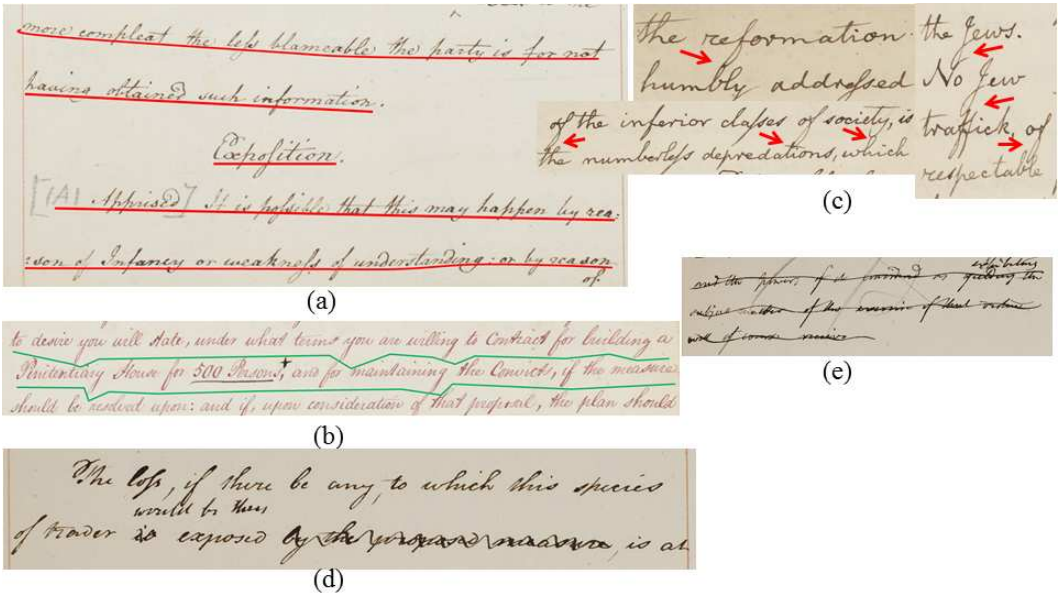


Figure 6. Challenges encountered on historical handwritten document images for text line segmentation. (a) Difference in the skew angle between lines on the page or even along the same text line, (b) overlapping text lines, (c) touching text lines, (d) additions above a text line, e) deleted text.

Smearing methods include the fuzzy run length smoothing algorithm (RLSA) (Shi and Govindaraju, 2004b), the adaptive local connectivity map method (Shi et al., 2005) and the proposal of Kennard and Barrett (2006). The fuzzy RLSA measure is calculated for every pixel on the initial image and describes "how far one can see when standing at a pixel along horizontal direction". By applying this measure, a new grayscale image is created which is binarized and the lines of text are extracted from the new image. The input to the adaptive local connectivity map method is a grayscale image (Shi et al., 2005). A new image is calculated by summing the intensities of each pixel's neighbors in the horizontal direction. Since the new image is also a grayscale image, a thresholding technique is applied and the connected components are grouped into location maps by using a grouping method. Kennard and Barrett (2006) presented a novel method for locating lines within free-form handwritten historical documents. Their method used an approach to find initial text line candidates which resembles the adaptive local connectivity map. The fuzzy RLSA as well as the adaptive local connectivity map method were evaluated using manuscripts written by Galileo, Newton and Washington showing a correct location rate of 93% and 95%, respectively. The method proposed by Kennard et al. was tested on 20 document images which were part of the Washington collection as well as 6 document images downloaded from the "Trails of Hope" showing encouraging performance.

Garz et al. (2012) proposed a text line segmentation method belonging to the grouping category which is binarization-free (input is a grayscale image), robust to noise and can cope with overlapping and touching text lines. First, interest points representing parts of characters are extracted from gray-scale images. At a next step, word clusters are identified

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

in high density regions and touching components such as ascenders and descenders are separated using seam carving. Finally, text lines are generated by concatenating neighboring word clusters, where neighborhood is defined by the prevailing orientation of the words in the document. Experiments conducted on the Latin manuscript images of the Saint Gall database (historical dataset) showed promising results for real-world applications in terms of both accuracy and efficiency. The work of Kleber et al. (2008) also belongs to the grouping category of methods. In this work, the authors presented an algorithm for ruling estimation of Glagolitic texts based on text line extraction which is suitable for degraded manuscripts by extrapolating the baselines with the a priori knowledge of the ruling. The algorithm was tested on 30 pages of the Missale Sinaiticum and the evaluation was based on visual criteria.

Hough based methods include the work of Louloudis et al. (2009). In this work, text line segmentation was achieved by applying the Hough transform on a subset of the document image connected components. A post-processing step included the correction of possible false alarms, the detection of text lines that the Hough transform failed to create and finally the efficient separation of vertically connected characters using a novel method based on skeletonization. The authors evaluated the method on a historical dataset of 40 images coming from the historical archive of the University of Athens as well as from the collection of George Washington using an established evaluation protocol which was first described in ICDAR 2007 Handwriting Segmentation Contest (Gatos et al., 2007). They reported an F-Measure of 99%. A hybrid method belonging to the Hough transform and grouping categories was proposed by Malleron et al. (2009). In this work, text line detection was modelled as an image segmentation problem by enhancing text line structure using Hough transform and a clustering of connected components in order to detect text line boundaries. Experiments showed that the proposed method can achieve high accuracy for detecting text lines in regular and semi-regular handwritten pages in the corpus of digitized Flaubert manuscripts.

Text line segmentation methods based on the seam carving principle were recently presented (Saabni et al. (2014), Arvanitopoulos and Süsstrunk (2014)). They try to segment text lines by finding an optimal path on the background of the document image travelling from the left to the right edge. Saabni et al. (2014) proposed a method which computes an energy map of a text image and determines the seams that pass across and between text lines. Two different algorithms were described (one for binary and one for grayscale images). Concerning the first algorithm (binary case), each seam passed on the middle and along a text line, and marked the components that make the letters and words of it. At a final step, the unmarked components were assigned to the closest text line. For the second algorithm (grayscale case) the seams were calculated on the distance transform of the grayscale image. Arvanitopoulos and Süsstrunk (2014) proposed an algorithm based on seam carving to compute separating seams between text lines. Seam carving is likely to produce seams that move through gaps between neighboring lines, if no information about the text geometry is incorporated into the problem. By constraining the optimization procedure inside the region between two consecutive text lines, robust separating seams can be produced that do not pass through word and line components. Extensive experimental evaluation on diverse manuscript pages showed improvement compared with the state-of-the-art for text line extraction in grayscale images.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 80 of 404

Other methodologies which cannot be grouped to a specific category include the works of Baechler et al. (2013), Chen et al. (2014) and Pastor-Pellicer et al. (2015). In more detail, Baechler proposed a text line extraction method for historical documents which works in two steps. In the first step, layout analysis is performed to recognize the physical structure of a given document using a classification technique. In the second step, the algorithm extracts the text lines starting from the layout recognition result. The system was evaluated on three historical datasets with a test set of 49 pages. The best obtained hit rate for text lines was 96.3%. Chen et al. (2014) used a pyramidal approach where at the first level, pixels are classified into: text, background, decoration, and out of page; at the second level, text regions are split into text line and non text areas. Finally, the text line segmentation results were refined by a smoothing post-processing procedure. The proposed algorithm was evaluated on three historical manuscript image datasets of diverse nature and achieved an average precision of 91% and recall of 84%. Finally, Pastor-Pellicer et al. (2015) proposed a text line extraction method with two contributions: first, supervised machine learning was used for the extraction of text-specific interest points; second, reformulating the problem of bottom-up text line aggregation as noise-robust combinatorial optimization. In a final step, unsupervised clustering eliminates invalid text lines. Experimental evaluation on the IAM Saint Gall historical dataset showed promising results.

Although a direct comparison of the abovementioned techniques cannot be made due to the fact that most methods use their own datasets and evaluation measures for measuring their method's performance, Table 5 briefly summarizes the size of the datasets as well the accuracy achieved by the methods just to give an idea on the performance of state-of-the-art methods.

Table 5. Comparison of performance for state-of-the-art text line segmentation methods.

Method	No. Document images	Evaluation Metric	Performance (%)
Bar-Yosef et al. (2009)	30	Visual	98
Shi and Govindaraju (2004b)	-	Manual	93
Shi et al. (2005)	30	Manual	95
Kennard and Barrett (2006)	26	Manual	-
Garz et al. (2012)	60	Line accuracy	97.97
Kleber et al. (2008)	30	Visual	95
Louloudis et al. (2009)	40	F-Measure	99
Saabni et al. (2014)	60	Correct Lines	98.9
Baechler et al. (2013)	14	Line accuracy	96.4
	30		95.4
	5		84.9
Pastor-Pellicer et al. (2015)	60	Line accuracy	97.2

3.3. Word Segmentation

Word segmentation refers to the process of defining the word regions of a text line. Since nowadays most handwriting recognition methods assume text lines as input, the word segmentation process is usually necessary only for segmentation-based query by example key-

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 81 of 404

word spotting methods. There are several challenges that need to be addressed by a word segmentation method (see Figure 7). These include the skew along a text line, the existence of slant angle among characters as well as punctuation marks which tend to reduce the inter word distance and the non uniform spacing of words.

Algorithms dealing with word segmentation in the literature are based primarily on the analysis of the geometric relationship between adjacent components. Related work for the problem of word segmentation differs in two aspects. The first aspect is the way the distance of adjacent components is calculated, while the second aspect concerns the approach used to classify the previously calculated distances as either between-word gaps or within-word gaps. Most of the methodologies described in the literature have a preprocessing stage which includes noise removal, skew and slant correction.

Many distance metrics are defined in the literature. Seni and Cohen (1994) presented eight different distance metrics. These include the bounding box distance, the minimum and average run-length distance, the Euclidean distance and different combinations of them which depend on several heuristics. Louloudis et al. (2009) proposed to use a combination of the Euclidean and the convex hull distance for the distance calculation stage, while using a novel gap classification method based on Gaussian mixture modeling. The authors report an F-Measure accuracy of 85.5% on a collection of 40 historical document images. It is assumed that the input of the word segmentation algorithm is the automatic text line segmentation result produced by their method.

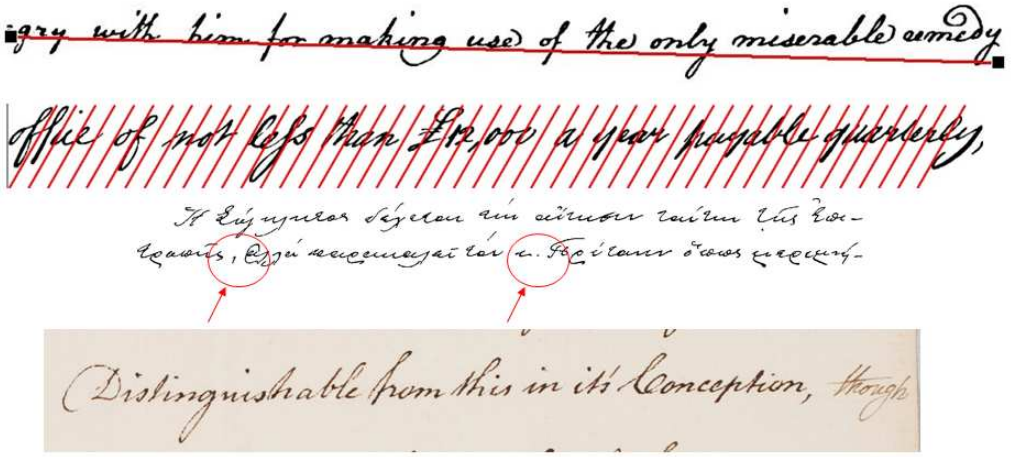


Figure 7. Challenges encountered on historical document images for word segmentation.

A different approach was proposed by Manmatha and Rothfeder (2005). In this work, a novel scale space algorithm for automatically segmenting handwritten (historical) documents into words was described. The first step concerns image cleaning, followed by a gray-level projection profile algorithm for finding lines in images. Each line image is then filtered with an anisotropic Laplacian at several scales. This procedure produces blobs which correspond to portions of characters at small scales and to words at larger scales. Crucial to the algorithm is scale selection, i.e. finding the optimum scale at which, blobs correspond to words. This is done by finding the maximum over scale of the extent or area of

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 82 of 404

the blobs. This scale maximum is estimated using three different approaches. The blobs recovered at the optimum scale are then bounded with a rectangular box to recover the words. A postprocessing filtering step is performed to eliminate boxes of unusual size which are unlikely to correspond to words. The approach was tested on a number of different data sets and it was shown that, on 100 sampled documents from the George Washington historical corpus of handwritten document images, a total error rate of 17% was observed. The technique outperformed a state-of-the-art word segmentation algorithm on this collection.

As it can be observed by the above mentioned descriptions, there is a lack of works dealing with the problem of word segmentation on historical documents. The main reason is related to the fact that recent methods for handwritten text recognition avoid the error-prone stage of word segmentation and thus start from the text lines in order to produce the final transcription. In addition, challenges which are met for the word segmentation step for historical document collections, do not exhibit large differences from the challenges encountered on modern collections. To this end, word segmentation methods developed for modern data may be used for the cases of historical data.

4. Handwritten Text Recognition (HTR)

Handwritten Text Recognition (HTR) becomes a challenging problem especially when dealing with historical documents. Major difficulties that appear concern (i) several degradations in the image quality, (ii) the large varieties in writing styles, language models, spelling rules and dictionaries, (iii) the use of abbreviations and special symbols as well as (iv) the limited amount of existing transcribed data that can be used for training. In this section, we assume that all necessary pre-processing and segmentation tasks have been already applied and the focus is on the pure recognition task. Based on the input that is provided to the recognition engine, we can distinguish the historical HTR methods to holistic and segmentation based. Holistic methods do not segment the image into characters but use as input the text line or the word image. On the other hand, segmentation-based approaches rely on segmentation into smaller entities which may correspond to characters or character parts. An overview of the HTR techniques for historical handwritten documents is given in Table 6

4.1. Holistic Methods for Recognition on Text Line Level

Two competitions have been organized for the recognition of historical handwritten documents starting from the corresponding text lines. These are the ICFHR-2014 HTRtS (Sanchez et al., 2014) and the ICDAR-2015 HTRtS (Sanchez et al., 2015) competitions. Both use manuscripts texts concerning legal reform, punishment, constitution, religion etc. written by the renowned English philosopher and reformer Jeremy Bentham (1748-1832) and his secretarial staff (see Figure 8a). The results are presented in Table 9 and show that when using more training data (Utrack) the word error rate can be less than 9% and the character error rate less than 3% using Multi-directional Long Short-Term Memory Neural Network (MDLSTM NN).

Bidirectional long short-term memory neural networks (BLSTM NN) have been used in Frinken et al. (2013) for the recognition of historical Spanish manuscripts. This work

Cerence Operating Company; Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 83 of 404

Table 6. Overview of HTR techniques proposed for historical handwritten documents

Reference	Input	Classifier	Features	Database	Evaluation results
Sanchez et al. (2014) (A2IA method)	Text line	MDLSTM NN	Gray scale image	ICFHR-2014 HTRtS competition (English)	Word error rate: 8.6% (Utrack) Character error rate: 2.9% (Utrack)
Sanchez et al. (2014) (CITlab method)	Text line	BPTT NN	Gray scale image		Word error rate: 14.6% (Rtrack) Character error rate: 5.0% (Rtrack)
Sanchez et al. (2014) (LIMSI method)	Text line	HMM-DNN HMM-LSTM NN	Handcrafted or pixel features from a sliding window		Word error rate: 15.0% (Rtrack) -11.0% (Utrack) Character error rate: 5.5% (Rtrack) - 3.9% (Utrack)
Toselli and Vidal (2015)	Text line	HMM	Sliding window - Geometric moments used for normalization		Word error rate: 18.5% (Rtrack) Character error rate: 7.5% (Rtrack)
Sanchez et al. (2015) (CITlab method)	Text line	BPTT NN	Gray scale image	ICDAR-2015 HTRtS competition (English)	Word error rate: 30.2% (Rtrack) Character error rate: 15.5% (Rtrack)
Sanchez et al. (2015) (A2IA method)	Text line	MDLSTM NN	Gray scale image		Word error rate: 31.6% (Rtrack) - 27.9% (Utrack) Character error rate: 14.7% (Rtrack) - 13.6% (Utrack)
Sanchez et al. (2015) (QCRI method)	Text line	HMM-DNN HMM-LSTM NN	Handcrafted or pixel features from a sliding window		Word error rate: 44.0% (Rtrack) Character error rate: 28.8% (Rtrack)
Frinken et al. (2013)	Text line	BLSTM NN	Sliding window - 9 geometric features	RODRIGO (Spanish)	Recognition Rate: 85.22%
Reese et al. (2014) (CIT System)	Word	MDRNN - CTC	Pixel data in four directions	ICFHR-2014 ANWRESH (English)	Total accuracy: 90.24% (age), 85.91 (birth place), 97.15% (marital status), 89.57% (relation), 70.08% (Given name), 49.69% (Surname)
Reese et al. (2014) (D1 System)	Word	K-NN	Zones - upper & lower profiles		Total accuracy: 45.92% (age), 40.75 (birth place), 90.73% (marital status), 79.40% (relation)
Reese et al. (2014) (D2 System)	Word	HMM	Sliding window - 9 geometric features		Total accuracy: 54.85% (age), 43.89 (birth place), 75.26% (marital status), 62.77% (relation)
Reese et al. (2014) (F1 System)	Word	CNN	Gray scale image		Total accuracy: 47.39 (birth place), 93.58% (marital status), 88.31% (relation)
Reese et al. (2014) (I2R System)	Word	RNN	Histogram of Oriented Gradients		Total accuracy: 72.90% (age), 47.24 (birth place), 95.72% (marital status), 91.30% (relation)
Lavrenko et al. (2004a)	Word	HMM	Fixed length feature vectors (e.g. word length, word profile)	George Washington (English)	Mean word error rate: 34.9%
Fischer et al. (2010)	Word	HMM	Graph similarity features	Parzival (German)	Word recognition accuracy: 94%
Fischer et al. (2009)	Word	BLSTM NN	Sliding window - 9 geometric features		Recognition rate: 93.32%
Ntzios et al. (2007)	Character	Binary Trees	Protrusions around cavities	Old Greek Early Christian	Average recall: 89.49% Precision: 98.06%
Saleem et al. (2014)	Character	NNDM	DSIFT	Missale Sinaiticum (Glagolitic)	Recall: 88.9% (normal characters), 70.8% (degraded characters)
Van Phan et al. (2016)	Character	k-d tree, GLVQ, MQDF2	Gradient features	Nom historical (Vietnamese)	Recognition rate: 66.92%
Tang et al. (2016)	Character	CNN	Convolutional layers are regarded as feature extractors	Dunhuang historical Chinese	Recognition accuracy: up to 70%

focuses on the language modelling aspect and demonstrates a recognition system that can cope with very large vocabularies of several hundred thousand words. It uses limited but accurate n-grams obtained from the training set and augment the language model with a very large vocabulary obtained from different sources. A sliding window is moved over the binary text line image to extract a sequence of 9 geometric features (Marti and Bunke, 2001): 3 global features which include the fraction of black pixels, the center of gravity and the second order moment as well as 6 local features which consist of the position of the upper and lower contour, the gradient of the upper and lower contour, the number of

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

black-white transitions and the fraction of black pixels between the contours. The database used in this work is the RODRIGO database which corresponds to a single-writer Spanish text written in 1545. Most of the pages consist of a single block of well-separated lines of calligraphical text (853 pages, 20356 lines) (see Figure 8b). The set of lines was divided into three different sets: training (10000 lines), validation (5010 lines), and test (5346 lines). The out-of-vocabulary rate of the test set is 6% given the vocabulary of the training and validation set. With the inclusion of external language sources, the out-of-vocabulary rate was significantly reduced from 6.15% to 2.80% (-3.35%) and by doing so, the recognition rate increased from 82.73% to 85.22% (+2.49%).

Traditional modelling approaches based on Hidden Markov optical character models (HMM) and an N-gram language model (LM) have been used in Toselli and Vidal (2015) for the recognition of the historical Bentham dataset used in the ICFHR-2014 HTRtS competition (Sanchez et al., 2014) (see Figure 8a). A set of 433 page images is used in this competition while 9198 text lines are used for training, 1415 for validation and 860 for testing. Departing from the very basic N-gram-HMM baseline system provided in HTRtS, several improvements are made in text image feature extraction, LM and HMM modelling, including more accurate HMM training by means of discriminative training. A narrow sliding window is horizontally applied to the line image for feature extraction. Geometric moments are used to perform some geometric normalizations to the images within each analysis window. The word error rate (WER) reported for the proposed system is 18.5% while the character error rate is 7.5%. These results are close to those achieved by deep and/or recurrent neural networks, including networks using BLSTM units.

4.2. Holistic Methods for Recognition on Word Level

A competition has been organized for the recognition of historical handwritten words. This is the ICFHR-2014 ANWRESH (Reese et al., 2014) that uses the ANWRESH dataset selected from the 1930 US Census collection including word bounding box and field lexicon data. In this competition, several teams submitted systems for recognizing six fields including Surname, Given Name(s), Age, Birth Place, Marital Status and Relation. The results are presented in Table 6 and show that a total accuracy of more than 90% can be achieve for closed-lexicon word recognition problems using multidimensional recurrent neural networks (MDRNN).

A holistic word recognition approach for single-author historical documents is presented in Lavrenko et al. (2004a). A HMM is used where words to be recognized represent hidden states. The state transition probabilities are estimated from word bigram frequencies. The observations are the feature representations of the word images in the document to be recognized. Feature vectors of fixed length are used ranging from coarse (e.g. word length) to more detailed descriptions (e.g. word profile). The evaluation corpus consists of a set of 20 pages from a collection of letters by George Washington (a total of 4856 words in the collection, 1187 of them unique) (see Figure 8c). A 20-fold cross-validation is carried out: during each iteration, one page is used as the testing page while the model is estimated from the remaining 19 pages. The proposed model achieved a mean word error rate of 35%, which corresponds to recognition accuracy of 65%.

Graph similarity features for historical handwritten word recognition based on HMMs

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy



Figure 8. Representative pages from (a) the Bentham, (b) the RODRIGO, (c) the George Washington and (d) the Parzival datasets.

is proposed in Fischer et al. (2010). The proposed graph similarity features rely on the idea of first transforming the image of a handwritten text into a large graph. Then local subgraphs are extracted using a sliding window that moves from left to right over the large graph. Each local subgraph is compared to a set of prototype graphs (each representing a letter from the alphabet) using a well-known graph distance measure. This process results in a vector consisting of n distance values for each local subgraph. Finally, the sequence of vectors obtained for the complete word image is input to a HMM-recognizer. The proposed method is tested on the medieval Parzival dataset (13th century). The manuscript is written in the Middle High German language with ink on parchment. Although several writers have contributed to the manuscript, the different writing styles are very similar (see Figure

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 86 of 404



Figure 9. Exemplary image portions of (a) Old Greek Early Christian manuscripts, (b) Glagolitic characters in the Missale Sinaiticum, (c) Nom scripts and (d) historical Chinese scripts.

8d). 11,743 word images are considered that contain 3,177 word classes and 87 characters including special characters that occur only once or twice in the dataset. The word images are divided into three distinct sets for training, validation, and testing. Half of the words is used for training and a quarter of the words for validation and testing, respectively. For each of the 74 characters present in the training set, a prototype graph is extracted from a manually selected template image. For five characters, two prototypes are chosen because two completely different writing styles were observed, resulting in a set of 79 prototypes. Consequently, the graph similarity features have a dimension of 79. A word recognition accuracy of 94% is reported.

Two state-of-the art recognizers originally developed for modern scripts are applied to medieval handwritten documents in Fischer et al. (2009). The first is based on HMMs and the second uses a Neural Network with a BLSTM architecture. Both word recognizers are based on 9 geometric features after applying a sliding window in the word image (Marti and Bunke, 2001). A Middle High German vocabulary is used without taking any language model into account. Each word is modelled by an HMM built from the trained letter HMMs and the most probable word is chosen using the Viterbi algorithm. For the NN based approach, the input layer contains one node for each of the nine geometrical fea-

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

tures and is connected with two distinct recurrent hidden layers. Both hidden layers are in turn connected to a single output layer. The network is bidirectional, i.e., the feature vector sequence is fed into the network in both the forward and the backward mode. The output layer contains one node for each possible letter in the sequence plus a special e node to indicate "no letter". For experimental evaluation, the Parzival database was used (45 pages of 4478 text lines). The set of all words is divided into a distinct training, validation, and test set. Half of the words is used for training and a quarter of the words for validation and testing, respectively. The NN-based recognizer with a BLSTM architecture outperformed the HMM-based recognizer with statistical significance (recognition rate: 93.32% for the NN-based recognizer, 88.69% for the HMM-based recognizer).

4.3. Recognition on Character Level

A method for detecting and recognizing character and character ligatures is presented in Ntzios et al. (2007) and is applied for the recognition of Old Greek Early Christian manuscripts. The continuity in writing for characters of the same or consecutive words as well as the unique characteristics of the lower case script in Early Greek Manuscripts (see Figure 9a) guided the authors to search for areas that contain open and closed cavities and then proceed to recognition by examining the topology of these areas and calculating the protrusions around them. The character recognition process consists of two basic stages. In the first stage each character is classified into a pattern based on the spatial configuration of cavities (e.g. the characters that have two vertical closed cavities are classified to the same pattern). In the second stage, for each pattern that corresponds to a unique character, there is a classification binary decision tree. Decision is taken at each node after the examination of specific feature value. For experimental evaluation, a dictionary of open and closed cavity patterns was built. A total of 12332 characters and character ligatures were used, from which 2497 characters were used for the training set and 9835 for the testing set. The proposed system recognizes basic characters with an average recall of 89.49% and precision of 98.06%.

The work of Saleem et al. (2014) deals with the recognition of Glagolitic characters in the Missale Sinaiticum written in the 11th century (see Figure 9b). An extension of the Dense SIFT (DSIFT) method is proposed in order to recognize Glagolitic characters. An image restoration is used as a preprocessing step to reduce background noise and enhance character strokes to improve the performance of DSIFT. At a next step, DSIFT features are computed in the test image and matched with the SIFT features of the restored training set images in order to localize and recognize Glagolitic characters using Nearest Neighbor Distance Maps (NNDM). Results using 15 image portions (913 normal and 142 degraded characters) show a Recall of 88.9% and 70.8% on normal and degraded characters.

The special case of Nom historical handwritten document recognition is considered in Van Phan et al. (2016) . Nom script (see Figure 9c) is the previous transcription system for vernacular Vietnamese language text widely used from the fifteenth to nineteenth centuries by Vietnam's cultured elite. According to this method, a character segmentation step splits the binarized images into individual character patterns. Then, the recognition step identifies class labels for character patterns automatically. The processing results can be

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

checked and corrected through a graphical user interface. The class labels of character patterns can be also fixed in the recognition step with another OCR version that can recognize an extended set of character categories. Finally, the documentation step completes the document recognition process by adding the character codes and layout information. The proposed character recognition system uses a k-d tree for coarse classification and the modified quadratic discriminant function (MQDF2) for fine classification. Training patterns were artificially generated from 27 Chinese, Japanese, and Nom character fonts since the three languages share a considerable number of character categories, and ground truth real patterns are not available for most Nom categories. Confining the character categories used for recognition in the first stage to the 7660 most frequently appearing categories increased the recognition rate to 66.92% from 55.50% for the extended set, which reduced the time and labour needed to manually tag unrecognized patterns.

A transfer learning method based on Convolutional Neural Network (CNN) is proposed in Tang et al. (2016) for historical Chinese character recognition. A CNN model is first trained by printed Chinese character samples. The network structure and weights of this model are used to initialize another CNN model, which is regarded as the feature extractor and classifier in the target domain. This is then fine-tuned by a few labelled historical or handwritten Chinese character samples, and used for final evaluation. The target domain includes 57,409 historical Chinese characters collected from Dunhuang historical Chinese documents (see Figure 9d). The results show that the recognition accuracy of the CNN based transfer learning method increases significantly as a number of samples for fine-tuning increases (70% when using up to 50 labelled samples of each character for fine-tuning).

5. Keyword Spotting

In cases where optical recognition is deemed to be very difficult or expected to give poor results, *word spotting* or *keyword spotting* has been proposed to substitute full text recognition. In word spotting the user queries the document database for a given word, and the spotting system is expected to return to the user a number of possible locations of the query in the original document. Keyword spotting has originally been proposed as an alternative for Automatic Speech Recognition (Rohlicek et al., 1989), as far back as 1989; in the mid-90's the first keyword spotting systems for document content began to appear (Manmatha and Croft, 1997).

The user may select the query by drawing a bounding box around the desired word in the scanned document image, or select the word from a collection of presegmented word images. This scenario is known as Query-by-example (QBE) in the literature. QBE keyword spotting is akin to content-based image retrieval (CBIR) (Sfikas et al., 2005). Both approaches follow the same paradigm, in the sense that the user defines an image query and the underlying system is required to detect matches of the query over a database. Features are extracted from the query and all database images, which are then used to build image descriptors. The descriptor of the query is then matched against the database images using a suitable distance metric. The matches that are found to be closest to the query are labelled as matches and are returned to the user.

The alternative to QBE is to expect from the user to type in as a string his query, in which

Cerence Operating Company, IPR2024-01267
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 89 of 404

case we have a Query-by-string (QBS) scenario. QBS also presupposes a pool of words, available as either a set of segmented words or segmented lines, for which the corresponding transcription is available. As in QBS we do not have image information about the query, the QBE/CBIR scheme of description building and matching cannot be applied.

The taxonomy of word spotting and recognition systems further includes the distinction into segmentation-based and segmentation-free systems. Segmentation-based methods assume that the scanned document image is segmented into layout elements up to the level of word. Segmentation-free approaches work with no such prior segmentation of the document; such approaches may be advantageous when the layout of the input image is too complex and the segmentation is expected to be of poor quality.

Machine learning methods have been employed in document understanding methods with much success, compared with the more standard learning-free approaches in document processing. Their basic assumption is that we can see word spotting and recognition as a (usually) supervised learning problem, where part of the data are expected to be labelled. Labelling in the document analysis context typically means that image data is beforehand related with a known alphanumeric transcription. In training time, the parameters of a suitable model are optimised using the labelled data. Learning-based methods in general are much more accurate than learning-free methods, even though their performance is dependent on the size and suitability of the training set compared to the test data. State-of-the-art learning-based models today include models based on Hidden Markov Models (HMM) and more recently models based on Neural Networks (NN) (España-Boquera et al., 2011; Frinken et al., 2010a,b).

5.1. State-of-the-Art Methods

In this subsection we shall attempt to review some of the most successful methods used in keyword spotting. Different keyword spotting methods use different assumptions about the query and data available (QBE vs QBS, segmentation-based vs segmentation-free), so we shall begin by examining the ”simplest” scenario, i.e. segmentation-based QBE, and examine other scenarios progressively.

Assuming that the query and the database are a collection of word images, some of the simplest features that have been proposed to describe a word image are column-based features or profile features (Rath and Manmatha, 2003). Profile features are defined as a set of scalar features per image column. In Rath and Manmatha (2003), after images are binarized, enhanced and skew/slant-normalized, features are extracted per column, including projection profiles, upper word profiles, lower word profiles and background/foreground transitions. Projection profiles are simply the sum of all foreground pixels per column. Upper(lower) word profiles record the distance of the word to the upper(lower) boundary. Background/foreground transitions record the number of transitions from a background to a foreground pixel, and vice versa, per column. Variations of this set of column-based features have been used elsewhere (Toselli and Vidal, 2013), adding different combinations of projections. Using a ”context” of neighbouring columns besides the central column of interest is also possible. As column-based features are by definition variable-length, Dynamic Time Warping (DTW) has been employed to match descriptors (Rath and Manmatha, 2003).

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 90 of 404

Zoning features have been used as an inexpensive way to build an efficient, fixed-length descriptor. In zoning features, the image is split in a fixed number of zones, forming a canonical grid over the image area. For each zone, a local descriptor is computed. In Sfikas et al. (2016), features extracted from a pre-trained Convolutional Neural Network (CNN) are computed per image zone, and then combined into a single, word-level fixed-length descriptor. Fixed-length descriptors have the advantage that they can be easily compared using an inexpensive distance such as the Euclidean or the Cosine distance (provided of course, that the comparison makes sense).

Since the beginning of the past decade, gradient-based features have succesfully been used for various computer vision applications (Dalal and Triggs, 2005; Lowe, 2004; Ahonen et al., 2006). Histograms of Gradients (HoG) (Dalal and Triggs, 2005) and Scale Invariant Feature Transform (SIFT) (Lowe, 2004) features describe an image area based on local gradient information. In the context of word image description, they can be used to efficiently encode stroke information locally. Gradient-based features are encoded into a single, word-level descriptor with an encoding / aggregation technique. In this direction, the Bag of Visual Words (BoVW) model has been employed (Aldavert et al., 2015). Input descriptors are used to learn a database-wide model that plays the role of a visual codebook, used subsequently to encode local descriptors of new images. Fisher Vectors (FV) extend on the BoVW paradigm, by learning a Gaussian Mixture Model (GMM) over the pool of gradient-based features, and using a measure of dissimilarity to the learned GMM to encode descriptors. FVs, coupled with SIFTs, have shown to lead to very powerful models for keyword spotting (Almazan et al., 2014b; Sfikas et al., 2015).

When word-level segmentation is not available, the word-to-word matching paradigm is evidently not applicable directly. Segmentation-free QBE can be useful when the scanned page is deemed too difficult to be segmented into word images correctly. One family of segmentation-free QBE approaches follows the approach of computing local keypoints on the unsegmented image. These keypoints are then matched with corresponding keypoints on the query image. In Leydier et al. (2007), an elastic matching method is used to match gradient-based keypoints. Heat kernel signature-based features are used as keypoints in Zhang and Tan (2013). Another approach to segmentation-free QBE spotting is to use a sliding window over the unsegmented image (Rothacker et al., 2013; Almazan et al., 2014a). As the process of matching a template versus a whole document image can be a computationally expensive process, assuming that a canonical grid of matching positions is used, heuristics have been proposed to bypass scanning the entire grid (Kovalchuk et al., 2014), or techniques to speed up matching, like product quantization (Almazan et al., 2014a).

In Almazan et al. (2014b) an attribute-based model has been proposed that uses supervised learning to learn its parameters. In this model, word attributes (Farhadi et al., 2009) are defined as a set of variates, each one of which corresponds to a certain word characteristic. This charasteristic may be the presence or absence of a certain character, bigram, or character diacritic (Sfikas et al., 2015). For each word image a vector of attributes is learned, encoding image information for each input. Attributes are then used together with ground-truth transcriptions to learn a projection from either attributes or transcriptions to a common latent subspace. It is worth noting that this model can be used to perform both QBE and QBS.

QBS keyword spotting for handwritten documents is performed typically with systems

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 91 of 404

that are based either on Hidden Markov Models (HMM) or the more recently used Recurrent Neural Networks (RNN). Both families of methods use supervised learning, so training with an annotated set is necessary before the models can be used in test mode, i.e. word spotting per se.

HMM models (Bishop, 2006) consist of two basic components. The one component is a series of hidden states that form a Markov chain. A finite number of possible hidden states is assigned to each possible character beforehand. These states are not directly observed from the data features, hence their characterisation as ”hidden”. Given each state, a distribution of observed features is defined. Emissions are typically modelled with a GMM (Bishop, 2006; Toselli and Vidal, 2013). Replacing GMM emissions with a standard feed-forward Neural Network has also shown good results (España-Boquera et al., 2011; Thomas et al., 2015). HMMs training is performed using the Baum-Welch algorithm to learn model parameters (Bishop, 2006). One HMM for each character is used to model the appearance of possible characters. Using a lexicon of possible words, the score of each word can be computed using the Viterbi algorithm for decoding (Bishop, 2006). Character HMMs can be used to also create a single HMM-”Filler” model, which can be used to decode inputs and detect words without a lexicon (Puigcerver et al., 2014).

HMM-based models were the state-of-the-art models for keyword spotting in handwritten documents (as well as for handwriting recognition systems), before the recent success of RNN-based models. Following the success of Fully-Connected and Convolutional Neural Networks in just about every field of Computer Vision, Recurrent NNs have shown to be well-suited especially for sequential data. Document data are modelled as sequences of column-based features. Text lines are typically used as both input and test data for RNNs, as is the case with the Bidirectional Long-Short term Memory Neural Network models (BLSTM-NN) (Frinken et al., 2012). BLSTM-NNs owe their name to a special part of their architecture, called *Long Short-term Memory Blocks* (LSTMs). LSTMs are used in order to mitigate the vanishing gradient problem (Frinken et al., 2012). The more recent Sequential Processing Recurrent Neural Networks (SPRNN) replace BLSTM-NN’s LSTM cells and bidirectional architecture with a different kind of architectural cell and Multidirectional / Multidimensional layers.

5.2. Evaluation of Keyword Spotting Methods

In keyword spotting research literature, new methods are usually evaluated by testing their efficiency on one or more historical manuscript collections. Some of the most used collections for spotting system evaluation, are the *Bentham* datasets and the *George Washington* dataset.

The *Bentham* dataset[1] was used in the H-KWS 2014 competition (Pratikakis et al., 2014). It contains 50 pages, handwritten by the English 18^{th} century utilitarian philosopher Jeremy Bentham and his secretaries. A second collection based on the writings of J. Bentham has been used, to which we shall refer here as the *Bentham-II*[2] dataset. It contains 70 pages segmented in $15,419$ segmented word images. It has been used as the testbed of the

[1] http://vc.ee.duth.gr/H-KWS2014/

[2] http://transcriptorium.eu/~icdar15kws/data.html

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

H-KWS 2015 competition (Puigcerver et al., 2015).

The *George Washington*[3] (GW) database (Lavrenko et al., 2004b) contains personal notes of the American 18^{th} century revolutionary. It contains 20 pages, 656 text lines and 4,894 words. While a significant number of works have presented results on GW, their results are usually not comparable to one another as many different evaluation protocols have been used for evaluation on GW (i.e. different evaluation queries and metrics).

A number of collections of historical documents written in non-latin scripts have also been used by the word spotting research community. These sets typically interest a more limited audience of researchers and scholars than their latin-script counterparts. For reference, we mention here the Arabic *Hadara* dataset (Pantke et al., 2014), containing 80 pages, written by the Palestinian El Hafid Ibn Hajr El Askalani in the 15^{th} century, and the Greek *Sophia Trikoupi* dataset[4], written in the 19^{th}*century* (Gatos et al., 2015). The two sets contain 80 and 46 pages respectively.

Precision at k (p@k) and Average Precision (AP) are arguably the two most widely used metrics, when one needs to quantify the performance of a keyword spotting system. Precision at k and AP are defined as

$$P@k = \frac{|\{\text{relevant instances}\} \cap \{\text{k retrieved instances}\}|}{k}$$

$$AP = \frac{\sum_{k=1}^{n} (P@k \times rel(k))}{\sum_{k=1}^{n} rel(k)}$$

where $rel(k)$ is an indicator function equal to 1 if the word at rank k is relevant and 0 otherwise. After a set of queries are defined, and the test system is used to retrieve matches for the set queries, metrics are evaluated per-query. Mean Average Precision (MAP), defined as the average value over all evaluation queries of the AP is then computed.

Table 7. Comparison of performance of *segmentation-based* keyword spotting methods on the *Bentham* database. Word-level segmentation was assumed to be available.

	%					
	Retsinas et al. (2016)	Sfikas et al. (2016)	Kovalchuk et al. (2014)	Almazan et al. (2014b)	Aldavert et al. (2015)	Howe (2013)
MAP	57.7	53.6	52.4	51.3	46.5	46.2
P@5	77.1	76.4	73.8	72.4	62.9	71.8

In tables 7 and 8 we show evaluation results for several recent keyword spotting methods. These methods are the NN-based Zoning Aggregated Hypercolumns (ZAH) (Sfikas et al., 2016), Attribute-based model (Almazan et al., 2014b), HOG/LBP-based method (Kovalchuk et al., 2014), Inkball model (Howe, 2013), Projections of Oriented Gradients (POG) (Retsinas et al., 2016), BoVW-based (Aldavert et al., 2015), elastic-matching model (Leydier et al., 2009) and template-matching model (Pantke et al., 2014). Some of these methods are available both as segmentation-based and segmentation-free methods (Kovalchuk et al.,

[3] http://www.iam.unibe.ch/fki/databases/iam-historical-document-database

[4] http://users.iit.demokritos.gr/~nstam/GRPOLY-DB/GRPOLY-DB-Handwritten.rar

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 93 of 404

Table 8. Comparison of performance of *segmentation-free* keyword spotting methods on the *Bentham* database.

	%			
	Kovalchuk et al. (2014)	Howe (2013)	Pantke et al. (2014)	Leydier et al. (2009)
MAP	41.6	36.3	33.7	20.5
P@5	60.9	55.6	54.3	33.5

2014; Howe, 2013). Also, we must note that the CNN-based ZAH model (Sfikas et al., 2016) and the attribute-based model (Almazan et al., 2014b) require a learning step; however, learning is assumed to be performed on a different set than the one that was used for testing ("*pre-training*"). ZAH is pre-trained on a large collection of street-view text, and the attribute-based model is pre-trained on the George Washington collection. All other methods do not require a learning phase. The best performance is given by the POG model (Retsinas et al., 2016) on the segmentation-based track, and the HOG/LBP-based model (Kovalchuk et al., 2014) on the segmentation-free track. It is worth noting that both winning methods rely on extracting gradient-based features. This fact validates the effectiveness of such features as descriptors of handwritten content.

Table 9. Comparison of performance of *learning-based* keyword spotting methods on the *Bentham-II* database. Results on a Query-by-String (QbS) and Query-by-Example (QbE) are shown.

	%			
	QbS		QbE	
	Strauß et al. (2016)	Puigcerver et al. (2015)	Strauß et al. (2016)	Puigcerver et al. (2015)
MAP	87.1	38.3	85.2	19.5
P@5	87.4	48.3	85.5	23.5

In table 9 we show numerical results that compare two state-of-the-art learning-based methods in keyword spotting. The two methods are: a Recurrent Neural Network (RNN) based method (Strauß et al., 2016), and a method based on the Hidden Markov Model (HMM) / filler model paradigm (Puigcerver et al., 2015). While the organizers of the related H-KWS 2015 competition state that the HMM-filler model used for the competition is a simple version and in general it can achieve better performance, we must note that the figures obtained for the NN-based model are impressive, in both QBS and QBE tracks. On the downside, the NN-based model requires a significant number of annotated text in order to be trained and subsequently used effectively, which in general may be non-trivial to obtain.

References

Afzal, M. Z., Pastor-Pellicer, J., Shafait, F., Breuel, T. M., Dengel, A., and Liwicki, M. (2015). Document image binarization using lstm: A sequence learning approach. In

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 94 of 404

Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing, HIP '15, pages 79–84, New York, NY, USA. ACM.

Ahonen, T., Hadid, A., and Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041.

Aldavert, D., Rusiñol, M., Toledo, R., and Llados, J. (2015). A study of bag-of-visual-words representations for handwritten keyword spotting. *International Journal on Document Analysis and Recognition*, 18(3):223–234.

Almazan, J., Gordo, A., Fornes, A., and Valveny, E. (2014a). Segmentation-free word spotting with exemplar SVMs. *Pattern Recognition*, 47(12):3967 – 3978.

Almazan, J., Gordo, A., Fornes, A., and Valveny, E. (2014b). Word spotting and recognition with embedded attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(12):2552–2566.

Arvanitopoulos, N. and Süsstrunk, S. (2014). Seam carving for text line extraction on color and grayscale historical manuscripts. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 726–731.

Asi, A., Cohen, R., Kedem, K., El-Sana, J., and Dinstein, I. (2014). A coarse-to-fine approach for layout analysis of ancient manuscripts. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 140–145.

Avidan, S. and Shamir, A. (2007). Seam carving for content-aware image resizing. *ACM Trans. Graph.*, 26(3).

Baechler, M. and Ingold, R. (2011). Multi resolution layout analysis of medieval manuscripts using dynamic mlp. In *Proceedings of the 2011 International Conference on Document Analysis and Recognition*, ICDAR '11, pages 1185–1189, Washington, DC, USA. IEEE Computer Society.

Baechler, M., Liwicki, M., and Ingold, R. (2013). Text line extraction using dmlp classifiers for historical manuscripts. In *Proceedings of the 2013 12th International Conference on Document Analysis and Recognition*, ICDAR '13, pages 1029–1033, Washington, DC, USA. IEEE Computer Society.

Baird, H. (2000). State of the art of document image degradation modeling. In *4th International Workshop on Document Analysis Systems (DAS) Invited talk*, pages 1–16. IAPR.

Bar-Yosef, I., Hagbi, N., Kedem, K., and Dinstein, I. (2009). Line segmentation for degraded handwritten historical documents. In *Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*, ICDAR '09, pages 1161–1165, Washington, DC, USA. IEEE Computer Society.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

Bukhari, S. S., Breuel, T. M., Asi, A., and El-Sana, J. (2012). Layout analysis for arabic historical document images using machine learning. In *Proceedings of the 2012 International Conference on Frontiers in Handwriting Recognition*, ICFHR '12, pages 639–644, Washington, DC, USA. IEEE Computer Society.

Bulacu, M., van Koert, R., Schomaker, L., and van der Zant, T. (2007). Layout analysis of handwritten historical documents for searching the archive of the cabinet of the dutch queen. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 1, pages 357–361.

Chen, K., Wei, H., Hennebert, J., Ingold, R., and Liwicki, M. (2014). Page segmentation for historical handwritten document images using color and texture features. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 488–493. IEEE.

Chen, Y. and Leedham, G. (2005). Decompose algorithm for thresholding degraded historical document images. *IEE Proceedings - Vision, Image and Signal Processing*, 152(6):702–714.

Cohen, R., Asi, A., Kedem, K., El-Sana, J., and Dinstein, I. (2013). Robust text and drawing segmentation algorithm for historical documents. In *Proceedings of the 2Nd International Workshop on Historical Document Imaging and Processing*, HIP '13, pages 110–117, New York, NY, USA. ACM.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In Schmid, C., Soatto, S., and Tomasi, C., editors, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 886–893.

Drira, F. (2006). Towards restoring historic documents degraded over time. In *Proceedings of the Second International Conference on Document Image Analysis for Libraries*, DIAL '06, pages 350–357, Washington, DC, USA. IEEE Computer Society.

España-Boquera, S., Castro-Bleda, M. J., Gorbe-Moya, J., and Zamora-Martinez, F. (2011). Improving offline handwritten text recognition with hybrid HMM/ANN models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):767–779.

Farhadi, A., Endres, I., Hoiem, D., and Forsyth, D. (2009). Describing objects by their attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1778–1785.

Farrahi Moghaddam, R. and Cheriet, M. (2010). A multi-scale framework for adaptive binarization of degraded document images. *Pattern Recogn.*, 43(6):2186–2198.

Fischer, A., Riesen, K., and Bunke, H. (2010). Graph similarity features for HMM-based handwriting recognition in historical documents. In *Proceedings of the* 12^{th} *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 253–258.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

Fischer, A., Wuthrich, M., Liwicki, M., Frinken, V., Bunke, H., Viehhauser, G., and Stolz, M. (2009). Automatic transcription of handwritten medieval documents. In *Proceedings of the* 15^{th} *International Conference on Virtual Systems and Multimedia (VSMM)*, pages 137–142.

Frinken, V., Fischer, A., and Bunke, H. (2010a). A novel word spotting algorithm using bidirectional long short-term memory neural networks. In *Proceedings of the* 4^{th} *Workshop on Artificial Neural Networks in Pattern Recognition*, volume 5998, pages 185–196.

Frinken, V., Fischer, A., Bunke, H., and Manmatha, R. (2010b). Adapting BLSTM neural network based keyword spotting trained on modern data to historical documents. In *Proceedings of the* 12^{th} *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 352–357, IEEE Computer Society, Washington, DC, USA.

Frinken, V., Fischer, A., Manmatha, R., and Bunke, H. (2012). A novel word spotting method based on recurrent neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2):211–224.

Frinken, V., Fischer, A., and Martínez-Hinarejos, C.-D. (2013). Handwriting recognition in historical documents using very large vocabularies. In *Proceedings of the* 2^{nd} *International Workshop on Historical Document Imaging and Processing (HIP2013)*, pages 66–72.

Gangamma, B., K, S. M., and Singh, A. V. (2012). Restoration of degraded historical document image. *Journal of Emerging Trends in Computing and Information Sciences*, pages 148–174.

Garz, A., Fischer, A., Sablatnig, R., and Bunke, H. (2012). Binarization-free text line segmentation for historical documents based on interest point clustering. In *Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on*, pages 95–99.

Gatos, B., Antonacopoulos, A., and Stamatopoulos, N. (2007). Handwriting segmentation contest. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02*, ICDAR '07, pages 1284–1288, Washington, DC, USA. IEEE Computer Society.

Gatos, B., Louloudis, G., and Stamatopoulos, N. (2014). Segmentation of historical handwritten documents into text zones and text lines. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 464–469.

Gatos, B., Ntirogiannis, K., and Pratikakis, I. (2009). Icdar 2009 document image binarization contest (dibco 2009). In *2009 10th International Conference on Document Analysis and Recognition*, pages 1375–1382.

Gatos, B., Pratikakis, I., and Perantonis, S. J. (2006). Adaptive degraded document image binarization. *Pattern Recogn.*, 39(3):317–327.

Gatos, B., Stamatopoulos, N., Louloudis, G., Sfikas, G., Retsinas, G., Papavassiliou, V., Sunistira, F., and Katsouros, V. (2015). Grpoly-db: An old greek polytonic document

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 97 of 404

image database. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 646–650. IEEE.

Hollaus, F., Diem, M., and Sablatnig, R. (2014). Improving ocr accuracy by applying enhancement techniques on multispectral images. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 3080–3085.

Howe, N. R. (2013). Part-structured inkball models for one-shot handwritten word spotting. In *Proceedings of the* 12^{th} *International Conference on Document Analysis and Recognition (ICDAR)*, pages 582–586.

Joo Kim, S., Deng, F., and Brown, M. S. (2011). Visual enhancement of old documents with hyperspectral imaging. *Pattern Recogn.*, 44(7):1461–1469.

Kennard, D. J. and Barrett, W. A. (2006). Separating lines of text in free-form handwritten historical documents. In *Second International Conference on Document Image Analysis for Libraries (DIAL'06)*, pages 12 pp.–23.

Kleber, F., Sablatnig, R., Gau, M., and Miklas, H. (2008). Ancient document analysis based on text line extraction. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4.

Kovalchuk, A., Wolf, L., and Dershowitz, N. (2014). A simple and fast word spotting method. In *Proceedings of the* 14^{th} *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 3–8.

Lavrenko, V., Rath, T., and Manmatha, R. (2004a). Holistic word recognition for handwritten historical documents. In *Proceedings of the Workshop on Document Image Analysis for Libraries (DIAL)*, pages 278–287.

Lavrenko, V., Rath, T. M., and Manmatha, R. (2004b). Holistic word recognition for handwritten historical documents. In *Proceedings of the* 1^{st} *International Workshop on Document Image Analysis for Libraries*, pages 278–287.

Leydier, Y., Bourgeois, F. L., and Emptoz, H. (2007). Text search for medieval manuscript images. *Pattern Recognition*, 40(12):3552– 3567.

Leydier, Y., Ouji, A., LeBourgeois, F., and Emptoz, H. (2009). Towards an omnilingual word retrieval system for ancient manuscripts. *Pattern Recognition*, 42(9):2089–2105.

Likforman-Sulem, L., Zahour, A., and Taconet, B. (2007). Text line segmentation of historical documents: a survey. *International Journal of Document Analysis and Recognition (IJDAR)*, 9(2):123–138.

Louloudis, G., Gatos, B., Pratikakis, I., and Halatsis, C. (2009). Text line and word segmentation of handwritten documents. *Pattern Recogn.*, 42(12):3169–3183.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

Lu, S., Su, B., and Tan, C. L. (2010). Document image binarization using background estimation and stroke edges. *International Journal on Document Analysis and Recognition (IJDAR)*, 13(4):303–314.

Malleron, V., Eglin, V., Emptoz, H., Dord-Crousl, S., and Rgnier, P. (2009). Text lines and snippets extraction for 19th century handwriting documents layout analysis. In *2009 10th International Conference on Document Analysis and Recognition*, pages 1001–1005.

Manmatha, R. and Croft, W. (1997). *Word spotting: indexing handwritten archives*, chapter 3, pages 43–64. MIT Press.

Manmatha, R. and Rothfeder, J. L. (2005). A scale space approach for automatically segmenting words from historical handwritten documents. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(8):1212–1225.

Marti, U. V. and Bunke, H. (2001). Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system. *Int. Journal of Pattern Recognition and Artificial Intelligence*, 15:65–90.

Moghaddam, R. F. and Cheriet, M. (2010). A variational approach to degraded document enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1347–1361.

Nicolas, S., Paquet, T., and Heutte, L. (2006). Complex handwritten page segmentation using contextual models. In *Second International Conference on Document Image Analysis for Libraries (DIAL'06)*, pages 12 pp.–59.

Ntirogiannis, K., Gatos, B., and Pratikakis, I. (2013). Performance evaluation methodology for historical document image binarization. *IEEE Transactions on Image Processing*, 22(2):595–609.

Ntirogiannis, K., Gatos, B., and Pratikakis, I. (2014a). A combined approach for the binarization of handwritten document images. *Pattern Recogn. Lett.*, 35:3–15.

Ntirogiannis, K., Gatos, B., and Pratikakis, I. (2014b). Icfhr2014 competition on handwritten document image binarization (h-dibco 2014). In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 809–813.

Ntzios, K., Gatos, B., Pratikakis, I., T., K., and S.J., P. (2007). An old greek handwritten OCR system based on an efficient segmentation-free approach. *International Journal on Document Analysis and Recognition*, 9:179–192.

Pantke, W., Dennhardt, M., Fecker, D., Märgner, V., and Fingscheidt, T. (2014). An historical handwritten arabic dataset for segmentation-free word spotting-hadara80p. In *14th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 15–20. IEEE.

Pastor-Pellicer, J., Garz, A., Ingold, R., and Castro-Bleda, M.-J. (2015). Combining learned script points and combinatorial optimization for text line extraction. In *Proceedings of*

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 99 of 404

the 3rd International Workshop on Historical Document Imaging and Processing, HIP '15, pages 71–78, New York, NY, USA. ACM.

Pratikakis, I., Gatos, B., and Ntirogiannis, K. (2010). H-dibco 2010 - handwritten document image binarization competition. In *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, pages 727–732.

Pratikakis, I., Gatos, B., and Ntirogiannis, K. (2011). Icdar 2011 document image binarization contest (dibco 2011). In *2011 International Conference on Document Analysis and Recognition*, pages 1506–1510.

Pratikakis, I., Gatos, B., and Ntirogiannis, K. (2012). Icfhr 2012 competition on handwritten document image binarization (h-dibco 2012). In *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*, pages 817–822.

Pratikakis, I., Gatos, B., and Ntirogiannis, K. (2013). Icdar 2013 document image binarization contest (dibco 2013). In *2013 12th International Conference on Document Analysis and Recognition*, pages 1471–1476.

Pratikakis, I., Zagoris, K., Gatos, B., Louloudis, G., and Stamatopoulos, N. (2014). ICFHR 2014 competition on handwritten keyword spotting (H-KWS 2014). In *Proceedings of the* 14^{th} *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 814–819.

Puigcerver, J., Toselli, A., and Vidal, E. (2014). Word-graph-based handwriting keyword spotting of out-of-vocabulary queries. In *Proceedings of the* 22^{nd} *International Conference on Pattern Recognition (ICPR)*, pages 2035–2040.

Puigcerver, J., Toselli, A., and Vidal, E. (2015). ICDAR2015 competition on keyword spotting for handwritten documents. In *Proceedings of the* 13^{th} *International Conference on Document Analysis and Recognition (ICDAR)*, pages 1176–1180.

Rath, T. M. and Manmatha, R. (2003). Word image matching using dynamic time warping. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 521–527.

Reese, J., Murdock, M., S., R., and Hamilton, B. (2014). ICFHR2014 competition on word recognition from historical documents (ANWRESH). In *Proceedings of the* 14^{th} *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 803–808.

Retsinas, G., Louloudis, G., Stamatopoulos, N., and Gatos, B. (2016). Keyword spotting in handwritten documents using projections of oriented gradients. In *Proceedings of the IAPR International Workshop on Document Analysis Systems (DAS)*, pages 411–416. IAPR.

Rohlicek, J., Russell, W., Roukos, S., and Gish, H. (1989). Continuous hidden Markov modeling for speaker-independent word spotting. In *Proceedings of the* 14^{th} *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 627–630 vol.1.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 100 of 404

Rothacker, L., Rusiñol, M., and Fink, G. A. (2013). Bag-of-features HMMs for segmentation-free word spotting in handwritten documents. In *Proceedings of the* 12^{th} *International Conference on Document Analysis and Recognition (ICDAR)*, pages 1305–1309.

Saabni, R., Asi, A., and El-Sana, J. (2014). Text line extraction for historical document images. *Pattern Recogn. Lett.*, 35:23–33.

Saleem, S., Hollaus, F., Diem, M., and Sablatnig, R. (2014). Recognizing glagolitic characters in degraded historical documents. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 771–776.

Sanchez, J. A., Romero, V., Toselli, A., and Vidal, E. (2014). ICFHR2014 competition on handwritten text recognition on transcriptorium datasets (HTRtS). In *Proceedings of the* 14^{th} *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 785–790.

Sanchez, J. A., Romero, V., Toselli, A., and Vidal, E. (2015). ICDAR 2015 competition htrts: Handwritten text recognition on the tranScriptorium dataset. In *Proceedings of the* 13^{th} *International Conference on Document Analysis and Recognition (ICDAR)*, pages 1166–1170.

Sauvola, J. and Pietikinen, M. (2000). Adaptive document image binarization. *Pattern Recognition*, 33(2):225 – 236.

Seni, G. and Cohen, E. (1994). External word segmentation of off-line handwritten text lines. *Pattern Recognition*, 27(1):41 – 52.

Sfikas, G., Constantinopoulos, C., Likas, A., and Galatsanos, N. P. (2005). An analytic distance metric for gaussian mixture models with application in image retrieval. In *International Conference on Artificial Neural Networks*, pages 835–840. Springer.

Sfikas, G., Giotis, A. P., Louloudis, G., and Gatos, B. (2015). Using attributes for word spotting and recognition in polytonic greek documents. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 686–690. IEEE.

Sfikas, G., Retsinas, G., and Gatos, B. (2016). Zoning aggregated hypercolumns for keyword spotting. In *15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, page "To appear". IEEE.

Shi, Z. and Govindaraju, V. (2004a). Historical document image enhancement using background light intensity normalization. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 1, pages 473–476 Vol.1.

Shi, Z. and Govindaraju, V. (2004b). Line separation for complex document images using fuzzy runlength. In *Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04)*, DIAL '04, pages 306–, Washington, DC, USA. IEEE Computer Society.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 101 of 404

Shi, Z., Setlur, S., and Govindaraju, V. (2005). Text extraction from gray scale historical document images using adaptive local connectivity map. In *Proceedings of the Eighth International Conference on Document Analysis and Recognition*, ICDAR '05, pages 794–798, Washington, DC, USA. IEEE Computer Society.

Strauß, T., Grüning, T., Leifert, G., and Labahn, R. (2016). Citlab ARGUS for keyword search in historical handwritten documents - description of citlab's system for the imageclef 2016 handwritten scanned document retrieval task. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.*, pages 399–412.

Su, B., Lu, S., and Tan, C. L. (2010). Binarization of historical document images using the local maximum and minimum. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, DAS '10, pages 159–166, New York, NY, USA. ACM.

Su, B., Lu, S., and Tan, C. L. (2013). Robust document image binarization technique for degraded document images. *IEEE Transactions on Image Processing*, 22(4):1408–1417.

Tan, C. L., Cao, R., and Shen, P. (2002). Restoration of archival documents using a wavelet technique. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(10):1399–1404.

Tang, Y., Peng, L., Xu, Q., Wang, Y., and A., F. (2016). CNN based transfer learning for historical chinese character recognition. In *Proceedings of the* 12^{th} *IAPR Workshop on Document Analysis Systems (DAS)*, pages 25–29.

Thomas, S., Chatelain, C., Heutte, L., Paquet, T., and Kessentini, Y. (2015). A deep hmm model for multiple keywords spotting in handwritten documents. *Pattern Analysis and Applications*, 18(4):1003–1015.

Tonazzini, A., Bedini, L., and Salerno, E. (2004). Independent component analysis for document restoration. *Document Analysis and Recognition*, 7(1):17–27.

Tonazzini, A., Salerno, E., and Bedini, L. (2007). Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique. *International Journal of Document Analysis and Recognition (IJDAR)*, 10(1):17–25.

Toselli, A. and Vidal, E. (2013). Fast HMM-Filler approach for key word spotting in handwritten documents. In *Proceedings of the* 12^{th} *International Conference on Document Analysis and Recognition (ICDAR)*, pages 501–505.

Toselli, A. H. and Vidal, E. (2015). Handwritten text recognition results on the bentham collection with improved classical N-Gram-HMM methods. In *Proceedings of the* 3^{rd} *International Workshop on Historical Document Imaging and Processing (HIP2015)*, pages 15–22.

Van Phan, T., Nguyen, K., and Nakagawa, M. (2016). A Nom historical document recognition system for digital archiving. *International Journal on Document Analysis and Recognition*, 19:49–64.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 102 of 404

Villegas, M. and Toselli, A. H. (2014). Bleed-through removal by learning a discriminative color channel. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 47–52.

Wolf, C. (2010). Document ink bleed-through removal with two hidden markov random fields and a single observation field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):431–447.

Zhang, X. and Tan, C. (2013). Segmentation-free keyword spotting for handwritten documents based on heat kernel signature. In *Proceedings of the* 12^{th} *International Conference on Document Analysis and Recognition (ICDAR)*, pages 827–831.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 103 of 404

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

In: Handwriting: Recognition, Development and Analysis ISBN: 978-1-53611-937-4
Editors: Byron L. D. Bezerra et al. © 2017 Nova Science Publishers, Inc.

Chapter 4

WAVELET DESCRIPTORS FOR HANDWRITTEN TEXT RECOGNITION IN HISTORICAL DOCUMENTS

Leticia M. Seijas[1,*] ***and Byron L. D. Bezerra***[2,†]
[1]Departamento de Computación, Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires, Buenos Aires, Argentina
[2]Escola Politécnica de Pernambuco
Universidade de Pernambuco (UPE), Recife, Brazil

1. Introduction

The automatic transcription of text in handwritten documents has many applications, from automatic document processing to indexing and document understanding. The automatic transcription of historical handwritten documents is an incipient research field that has been started to be explored in recent years. For some time in the past decades, the interest in Off-line Handwritten Text Recognition (HTR) was diminishing under the assumption that modern computer technologies will soon make paper-based documents useless. However, the increasing number of on-line digital libraries publishing large quantities of digitized legacy papers and the fact that the transcription of them into a textual electronic format would provide historians and other researches new ways of indexing and easy retrieval, have turned HTR up in a major research topic (Sánchez et al., 2014).

HTR for historical documents is a high complex task mainly because of the strong variability of writing styles, different font types and sizes of characters, underlined and/or crossed-out words. Moreover, this complexity is increased by the presence of typical degradation problems of ancient documents such as background variability and the presence of spots due to the humidity or marks resulting from the ink that goes through the paper. For this reason, different methods and techniques of document analysis and recognition fields are needed.

The nowadays common technology for HTR is based on a segmentation-free approach, where the recognition system is able to recognize all the text elements (sentences, words,

*E-mail address: lseijas@dc.uba.ar.
†E-mail address: byronleite@ecomp.poli.br.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

and characters) as a whole, without any prior segmentation of the image into these elements (Sánchez et al., 2014; Marti and Bunke, 2002; Toselli et al., 2004a; Espana-Boquera et al., 2011).

The use of segmentation-free (holistic) techniques which tightly integrate an optical character model and a language model has yielded the best performance on standard benchmarks. Although the N-grams language models and the Gaussian Mixture Hidden Markov Models (HMM-GMM) have been considered the most traditional and better-understood approaches in the last years, recently some Artificial Neural Networks (ANN) have gained considerable popularity in the HTR research community (Toselli and Vidal, 2015; Gouveia et al., 2014; Bezerra et al., 2012). A large amount of research has been done to improve these recognition models and to develop the corresponding training and decoding algorithms (Bluche, 2015).

On the other hand, feature extraction strategies have not been widely explored. In the segmentation-free method, the preprocessed line image is segmented into frames using a sliding window to extract features from each slice. Some of the techniques for feature extraction presented in previous works are based on the computation of pixel densities, raw gray levels and their gradients, geometric moment normalization, Principal Components Analysis (PCA) to reduce and decorrelate the pixel dimensions.

Menasri et al. (2011) presented an efficient word recognition system resulting from the combination of three handwriting recognizers. The main component of this combined system is a HMM-based recognizer which considers dynamic and contextual information for a better modeling of writing units. Neural networks (NN) are also used. Feature extraction is based on the work of Mohamad et al. (2009); El-Hajj et al. (2005). Using the segmentation-free approach, the windows are divided vertically into a fixed number of cells. Within each window, a set of geometric features is extracted: w features are related to pixel densities within each window's column (w is the width of the extraction window, in pixels). There are three density features extracted from the whole frame and above and under the lower baseline. Two features are related to background/foreground transitions between adjacent cells. Three features are for the gravity center position, including a derivative feature (difference between y positions). Twelve other features are related to local pixel configurations to capture stroke concavities (Menasri et al., 2011). A subset of features is baseline dependent. The final descriptor has 28 components.

Michal et al. (2013); Kozielski et al. (2012) proposed a HMM-based system for off-line handwriting recognition based on successful techniques from the domains of large vocabulary speech recognition and image object recognition. This work introduces a moment-based scheme for preprocessing and feature extraction. The preprocessing stage includes normalizing the contrast of the gray-scale image and fixing the slant on the text line segmented from the text pages. Then, frames are extracted with and overlapping sliding window and the 1st- and 2nd-order moments are calculated for each frame independently. The 1st-order moments represent the center of gravity which is used to shift the content of the frame to the center of the image. The 2nd-order moments correspond to the weighted standard deviation of the distance between pixels in the frame and the center of gravity. They are used to compute the scaling factors for size and translation normalization. This way, every frame extracted with the sliding window is normalize using the scaling factors. Then, the gray-scale values of all pixels in a normalized frame are used as features and are fur-

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 106 of 404

ther reduced by PCA to 20 components. During normalization, the aspect ratio is not kept because the vertical and horizontal moments are computed and normalized separately. For this reason, four values related to the original moments are added in order to map the class specific moment information, which was originally distributed over the whole image, to specific components of the feature vector. The final feature vector has 24 dimensions.

Toselli and Vidal (2015) presented HTR experiments and results on the historical Bentham text image dataset used in the ICFHR-2014 HTRtS competition, adopting the segmentation-free holistic framework and using traditional modeling approaches based on Hidden Markov optical character models (HMM) and an N-gram language model (LM). Departing from the very basic N-gram-HMM baseline system provided in HTRtS, several improvements were made in LM and HMM modeling. It includes more accurate HMM training through discriminative training, achieving similar recognition accuracy as some of the best performing (single, uncombined) systems based on (recurrent) Neural Networks, using identical training and testing data. For feature extraction, a narrow sliding window was horizontally applied to the preprocessed line image. For each window position i; $1 \leq i \leq n$, a 60-dimensional feature vector was obtained by computing three kinds of features: normalized vertical gray level s for 20 evenly distributed vertical positions and horizontal and vertical gray level derivatives at each of these vertical positions. The features proposed in Michal et al. (2013) were also used.

In the thesis work of Bluche (Bluche, 2015), a study of different aspects of optical models based on Deep Neural Networks in a hybrid Neural Network / HMM scheme was conducted, to better understand and evaluate their relative importance. First, it is showed that Deep Neural Networks produce consistent and significant improvements in networks with one or two hidden layers, independently of the kind of neural network, MLP or RNN, and of input, handcrafted features or pixels. Despite the dominance of LSTM-RNNs in the recent literature on handwriting recognition, it can be seen that deep MLPs achieve comparable results. This work also evaluated different training criteria reporting similar improvements for MLP/HMMs as those observed in speech recognition, with sequence-discriminative training. The proposed approach was validated by taking part in the HTRtS contest in 2014. For feature extraction, the sliding window framework is applied. Two kinds of features are extracted from each window: handcrafted and pixel intensities. The first ones are geometrical and statistical features used in the work of Menasri et al. (2011), that result in a descriptor of size 56. The "pixel features" are calculated from a downscale frame with its aspect ratio kept constant, and then transformed to lie in the interval [0,1]. An 800-dimensional feature vector is obtained for the Bentham set. The results of the final systems presented in this thesis, namely MLPs and RNNs, with handcrafted feature or pixel inputs, are comparable to the state-of-the-art on Rimes and IAM datasets. The combination of these systems outperformed all published results on the considered databases.

This work proposes a different approach for feature extraction based on the application of the CDF 9/7 Wavelet Transform for the HTR problem. The wavelet transform has been applied to related areas such as handwritten character recognition (Patel et al., 2012; Seijas and Segura, 2012) and speech recognition (Trivedi et al., 2011; Shao and Chang, 2005). Our approach improves data representation as a result of considerably reducing the feature vector size while retaining the basic structure of the pattern, and provides competitive HTR results. Section "HTR Systems Based on HMM/GMM" gives an overview of

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 107 of 404

HTR systems based on the HMM-GMM model. In Subsection "The Wavelet Transform" fundamentals of the Discrete Wavelet Transform are introduced while in Subsection "The proposed WT-descriptors" the wavelet-based descriptors are proposed. Section "Experiments and Results" reports the experiments and results and finally, the conclusions of the work are presented in Section "Conclusion".

2. HTR Systems Based on HMM/GMM

Handwriting recognition consists of several steps, from the preparation of the image to the delivery of the recognized character or word sequences. Generally, the inputs to the recognition systems are images of word or text lines, which must sometimes be extracted from document images. Image processing techniques attempt to reduce the variability of writing style, including those that normalize the image quality the size of the writing (Bluche, 2015). The extraction of relevant features from the image also eliminates some of the diversity, and aims at producing pertinent values which represent the problem in a data space where it is easier to solve.

The most traditional and better-understood modeling approaches for HTR are N-grams for the language models and Gaussian Mixture Hidden Markov Models (HMM-GMMs) for the optical models. The traditional N-gram/HMM-GMM framework offers several advantages over modern approaches based on (hybrid, recurrent) NNs. Perhaps the most important are the much faster training of HMMs and the well-understood stability of the results of Baum-Welch training. These advantages become crucial when dealing with many historical document collections, which are typically huge and entail very high degrees of variability, making it often difficult to re-use models trained on previous collections (Toselli and Vidal, 2015).

The fundamentals of HTR based on N-gram/HMM-GMM were originally presented in Bazzi et al. (1999) and further developed in Bazzi et al. (1999); Toselli et al. (2004a), among others. Recognizers of this kind accept an input text line image, represented as a sequence of feature vectors $x = \{\overrightarrow{x_1}, \overrightarrow{x_2}, \cdots, \overrightarrow{x_n}\}, \overrightarrow{x_i} \in \Re^D$ and find a most likely word sequence $w = w_1 w_2 \cdots w_l$, according to:

$$\widehat{w} = \arg\max_w P(w|x) = \arg\max_w P(W)\, p(w|x) \tag{1}$$

The prior probability $P(w)$ is approximated by an N-gram LM, and the conditional density $p(x|w)$ is approximated by combining (generally just concatenating) the character HMMs of the words in w. Each character (alphabet element) is modeled by a continuous density left-to-right HMM, where a Gaussian mixture model (GMM) is used to account for the emission of feature vectors in each HMM state. Once a HMM topology (number of states, Gaussians, and structure) has been adopted, the model parameters can be easily estimated by maximum likelihood. The required training data consists of continuous handwriting text line images (without any word or character segmentation), accompanied by the transcription of this text into the corresponding sequence of characters. This training process is carried out using a well-known instance of the EM algorithm called *embedded Baum-Welch* re-estimation (Jelinek, 1999). Maximum-likelihood parameter estimation is

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 108 of 404

the simplest, most basic training approach for HMMs (Toselli and Vidal, 2015). Figure 1 shows a prototype of HMM with left-to-right topology having six states.

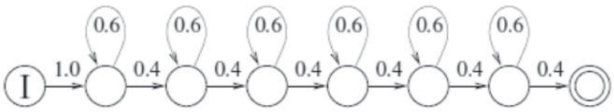


Figure 1. A prototype of HMM topology having 6 states.

The decoding problem in Equation 1 can be solved by the Viterbi algorithm (Jelinek, 1999). Figure 2 depicts this decoding process and the models involved. More details can be found in Romero et al. (2012); Young et al. (2009).

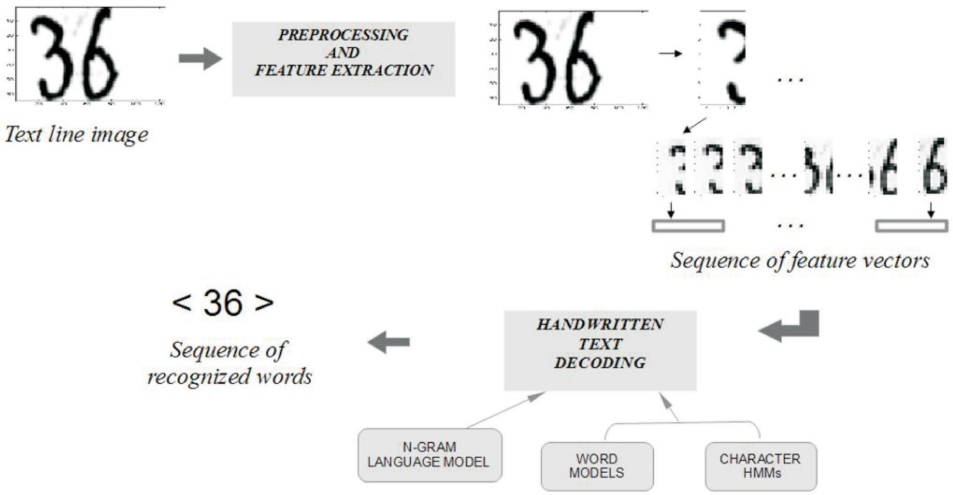


Figure 2. HTR decoding. For a text line image that could include different characters (in the example a handwritten number “36”), a feature vector sequence is produced. Then, this sequence is decoded into a word sequence using three knowledge sources: optical character HMMs, a lexicon (word models) and a language model (Toselli and Vidal, 2015).

3. Wavelet-Based Descriptors for HTR

We considered text line images as the basic input of the HTR process. They can be obtained from each document image using conventional text line detection and segmentation techniques (Likforman-Sulem et al., 2006; Bosch et al., 2012; Toselli and Vidal, 2015). The extracted lines are preprocessed to clean and enhance images, to correct skewed lines and slanted handwriting and normalized the size of images (Pastor et al., 2004a, 2006). Then, a sliding window is horizontally applied to the preprocessed line. For each frame extracted (corresponding to each window), a feature vector is obtained. In this segmentation-free approach, the recognition is accomplished without an explicit segmentation of the image, thus

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 109 of 404

without relying on heuristics to find character boundaries, and limiting the risk of under-segmentation. This category of approaches is the most popular nowadays, receiving a lot of research interest, and achieving the best performance on standard benchmarks (Bluche, 2015). Algorithm 1 shows the basic steps of our feature extraction proposal.

Algorithm 1 Wavelet-based feature extraction proposal for HTR

```
for all preprocessed line image do
    repeat
        Extract a frame using the sliding window approach;
        Apply the WT to the frame
        Apply PCA transformation to the LL subband of the WT calculated in previous
        step to obtain the descriptor.
    until All the frames in a line are processed.
end for
```

Subsection "Experimental Setup" describes this process with values extracted from experiments.

3.1. The Wavelet Transform

The Wavelet Transform (WT) is a technique particularly suited for locating spatial and frequential information in image processing and, particularly, for feature extraction from patterns to be classified. Many works have applied WT in different areas (Pastor et al., 2004b; Chen et al., 2006), including handwritten digit recognition (Seijas and Segura, 2012).

The Discrete Wavelet Transform (DWT) is based on the subband-coding technique, being a variant easy to implement of WT that requires a few resources and computing time. The DWT is well suited for multiresolution analysis (MRA) and lossless reconstruction by the use of filter banks (Debnath, 2002).

The Fast Orthogonal Wavelet Transform (FWT) decomposes each approximatioP_{V_j} f n of a function $f \in L^2\ (\Re)$ into approximations of lower resolution $P_{V_{j+1}}$ f plus wavelet coefficients produced by the projection $P_{W_{j+1}}$ f, being V_j a multiresolution approximation (S. Mallat, 1999), W_j the orthogonal complement of V_j and P_{V_j} f the orthogonal projection of f in $V_j, j \in Z$. Conversely, for the reconstruction from wavelet coefficients, each P_{V_j} f is obtained through $P_{V_{j+1}}$ f and $P_{W_{j+1}}$ f. Since $\{\varphi_{j,n}; j,n \in Z\}$ and $\{\Psi_{j,n}; j,n \in Z\}$ are orthonormal bases for V_j and W_j with φ and Ψ being scale and wavelet functions respectively, the projections on these subspaces are defined by:

$$a_j = \langle f, \varphi_{j,n} \rangle, d_j[n] = \langle f, \Psi_{j,n} \rangle \tag{2}$$

In Equation 2, $a_j[n]$ represents the approximation coefficients, $d_j[n]$ the detail ones and $\langle \cdot, \cdot \rangle$ is the inner product operation. The Mallat algorithm (S. Mallat, 1999) allows computing the coefficients through convolutions and subsamples in cascade:

$$a_{j+1}[p] = \sum_{n=-\infty}^{\infty} h[n-2p] a_j[n] = a_j \bar{h}[2p] \tag{3}$$

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 110 of 404

$$d_{j+1}[p] = \sum_{n=-\infty}^{\infty} g[n-2p]a_j[n] = a_j\bar{g}[2p] \tag{4}$$

where $\bar{x}[n] = x[-n]$ and h,g are high and low frequency filters, respectively. Equations 3 and 4 correspond to the decomposition stage (see Figure3). In each level, the high-pass filter generates detail information d_j while the low-pass filter, associated with the scale function, produces approximations, smoothed representations a_j of the signal. Implementing the FWT requires just $O(N)$ operations for signals of size N providing a non-redundant representation and allowing for lossless reconstruction. The filter bank algorithm with perfect reconstruction can also be applied to the Fast Biorthogonal Wavelet Transform. The wavelet coefficients are computed by successive convolutions with filters h and g while for reconstruction the dual filters $\tilde{h}$ and $\tilde{g}$ are used. If the signal consists of N non-zero samples, to compute its representation based on the biorthogonal wavelet, $O(N)$ operations are needed (S. Mallat, 1999).

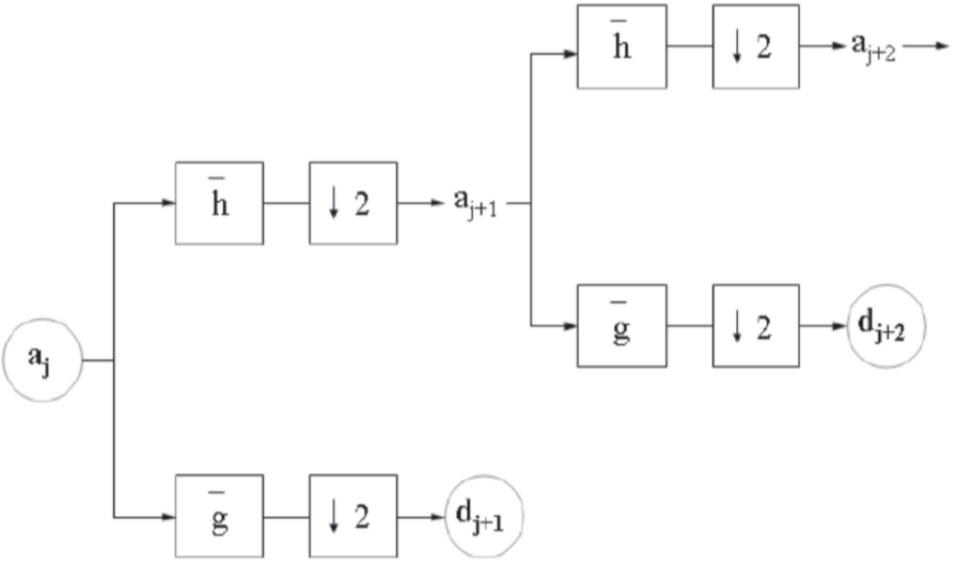


Figure 3. FWT decomposition using h and g filters and downsampling ($\downarrow 2$).

So far we have dealt with the DWT in one dimension. Digital image processing requires a bidimensional WT. It is computed by application of the one-dimensional FWT first to rows and then to columns. Let $\Psi(x)$ be the one-dimensional wavelet associated to the one-dimensional scale function $\varphi(x)$, then the scale function in two dimensions is:

$$\varphi(x,y)_{LL} = \varphi(x)\varphi(y) \tag{5}$$

The three bidimensional wavelets are defined by:

$$\Psi(x,y)_{LH} = \varphi(x)\psi(y) \tag{6}$$

$$\Psi(x,y)_{HL} = \psi(x)\varphi(y) \tag{7}$$

$$\Psi(x,y)_{HH} = \psi(x)\psi(y) \tag{8}$$

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 111 of 404

where *LL* represents lowest frequencies (global information)), *LH* represents high vertical frequencies (horizontal details), *HL* high horizontal frequencies (vertical details) and *HH* high frequencies on both diagonals (diagonal details).

Application of a step of the transform on the original image produces an approximation subband *LL* corresponding to the smoothed image, and three detail subbands *HL*, *LH* and *HH*. The following step works on the approximation subband, resulting in four subbands as can be seen in Figure 4. In other words, each step of the decomposition represents the approximation subband of level *i* as four subbands at level *i+1*, each one being a quarter in size respecting the original subband.

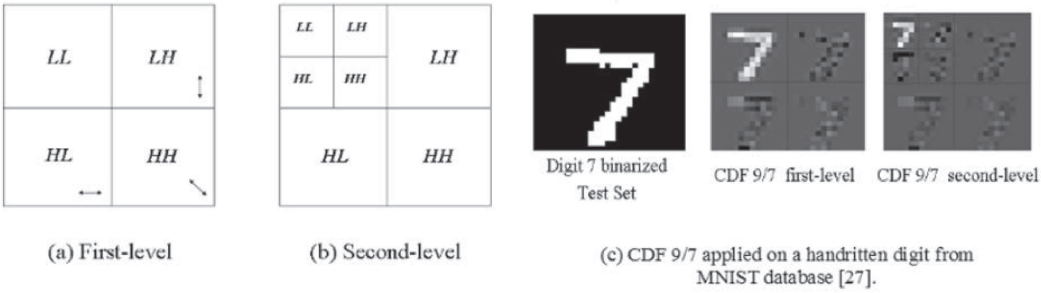


Figure 4. Multilevel decomposition of an N x N image using a 2D-DWT (Seijas and Segura, 2012).

The Biorthogonal Bidimensional Wavelet Cohen-Daubechies-Feauveau (CDF) 9/7 was efficiently applied to the JPEG2000 compression standard, and also in fingerprint compression by the FBI (Skodras et al., 2001). It was also applied for pattern representation in classification processes. In Seijas and Segura (2012) descriptors for handwritten numeral recognition based on multiresolution features by the use of the CDF 9/7 Wavelet Transform and Principal Component Analysis (WT-PCA) were proposed, improving classification performance with a considerable reduction (near the 90%) of the dimensionality of the representation.

Figure 5 shows the scale and wavelet functions and the coefficients of the corresponding filters, for the CDF 9/7 transform in the decomposition.

3.2. The Proposed WT-Descriptors

Within the segmentation-free framework, we applied the CDF 9/7 wavelet transform to each frame extracted from the gray-scale line image. Different combinations were evaluated for constructing the descriptor or feature vector, considering the subbands obtained at different levels of resolution with the WT and including the thresholding of the coefficients which sometimes improves image quality reducing noise (Dewangan and Goswami., 2012). The approximation subbands (*LL*) produce smoothed images of the pattern, preserving shape and reducing the dimension to a quarter of the original size at a first level, 16th in a second level, and 2^{2*l} at level *l* where the image is coarser. The high-frequency subband HH shows sudden changes in image contours (diagonal details), and also the LH and HL subbands providing vertical and horizontal features on the smoothed pattern (see Figure 4.c).

Declaration of Dr. Byron Leite Dantas Bezerra, 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 112 of 404

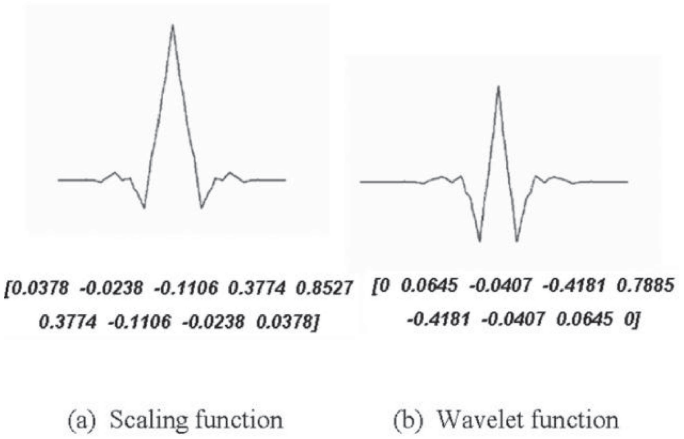


Figure 5. MCDF 9/7 with filters for signal decomposition (Seijas and Segura, 2012).

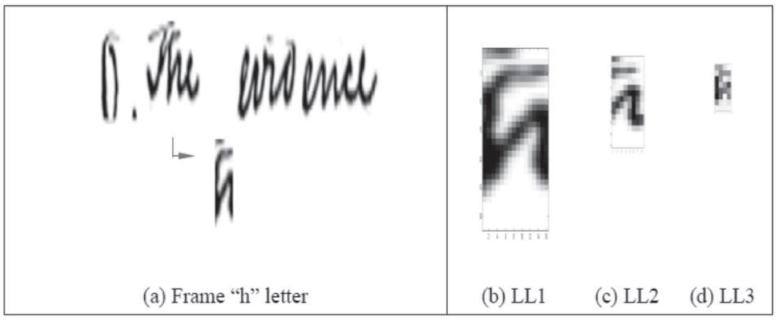


Figure 6. Approximation subbands of CDF 9/7 at different levels of resolution.(a) a frame ("h" letter) of size 64x32 pixels extracted from a preprocessed text line of Bentham database; (b) approximation subband *LL* at level 1, size 32x16; (c) *LL* at level 2, size 16x8; (d) *LL* at level 3, size 8x4.

After several preliminary experiments, we concluded the detail coefficients did not contribute to improving HTR rates. Therefore, descriptors using only the approximation subbands from level 1 to 3 with non-thresholded representation were finally considered because of the best results obtained. We think that this approach retains the basic structure of the pattern, eliminating details that do not improve the classification. The feature vector is subject to PCA transformation to reduce the size of the descriptor using the directions that contain most of the data variance while disregarding those with little information.

The following are the descriptors selected:

A1: approximation subband of level 1 (LL1) + PCA.
A2: approximation subband of level 2 (LL2) + PCA.
A3: approximation subband of level 3 (LL3) + PCA.

In Figure 6 an example of approximation subbands at different levels of resolution is presented. It can be seen that while the frame extracted from the text line has $64x32 = 2048$

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 113 of 404

gray-scale values, the LL subband at level 1 (LL1) has *32x16=512* wavelet coefficients, reducing the size of the representation in a quarter and retaining the structure of the pattern. LL at level 2 (LL2) has *16x8=128* coefficients and LL at level 3 (LL3) has *8x4=32* values. With this technique, we achieve a strong reduction of the pattern representation in size, while the image becomes coarser, retaining basic shapes.

4. Experiments and Results

4.1. Bentham Database

The whole set contains more than 80,000 images of manuscripts written by the renowned English philosopher and reformer Jeremy Bentham (1748-1832) and his secretarial staff (Causer and Wallace, 2012). It includes texts about legal reform, punishment, the Constitution, and religion. The data were prepared by University College, London, during the TRANSCRIPTORIUM project (Sánchez et al., 2013). The transcription of this collection is currently being carried out by volunteers participating in the award-winning crowdsourcing initiative known as "Transcribe Bentham" (Toselli and Vidal, 2015). Page images of the Bentham collection (see Figure 7) usually entail several difficulties for the complete recognition process, due to the presence of marginal notes, fainted writing, stamps, skewed images, slanted lines, different writing styles, crossed-out lines, hyphenated words, punctuation symbols, among others. Even with these difficulties, most of these page images are readable for human beings (Sanchez et al., 2015).

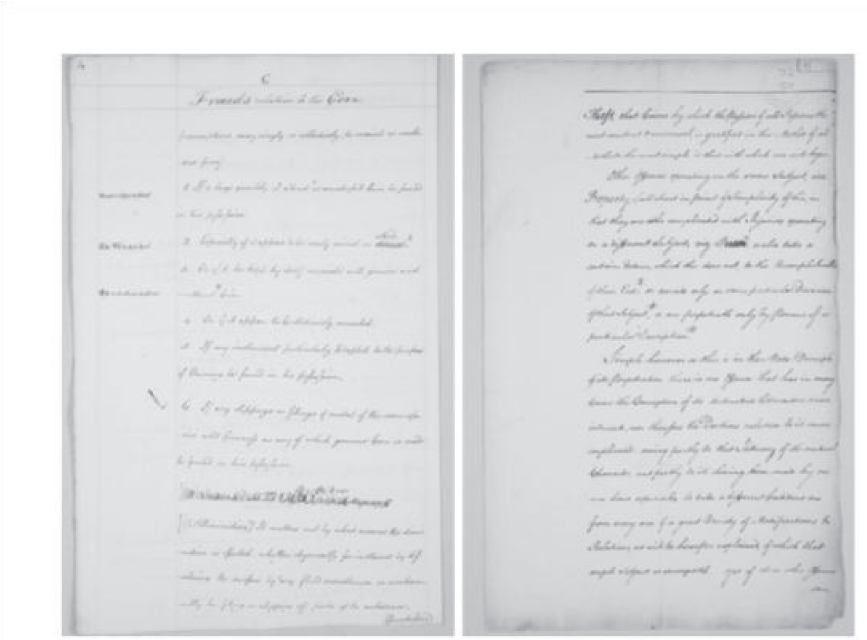


Figure 7. Examples of Bentham page images (Toselli and Vidal, 2015).

For the experiments, we chose the particular data set and partitions used in the ICFHR

Cerence Operating Company Exhibit 2012
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 114 of 404

2014 HTRtS competition (Causer and Wallace, 2012) described in Table 1. Figure 8 shows sample lines extracted from the page images and provided for experimentation. A total of 11473 lines were used, 10613 for training and tuning the system and 860 for testing. The training set was divided into two subsets of an equal number of lines to evaluate and select the feature vectors with better performance. Final percentages were obtained by training the recognition system using the training and validation set (see Table 1) and testing with the 860 test lines as suggested in the ICFHR 2014 HTRtS Restricted Track.

Table 1. Bentham dataset used in ICHFR-2014 HTRtS contest (Causer and Wallace, 2012; Toselli and Vidal, 2015)

Number of:	Training	Validation	Test	Total
Pages	350	50	33	433
Lines	9198	1415	860	11473
Running words	86075	12962	7868	106905
Lexicon	8658	2709	1946	9716
Running OOV*	-	857	417	-
OOV Lexicon	-	681	377	-
Char. set size	86	86	86	86
Run.Characters	442336	67400	40938	550674

*OOV: out of vocabulary

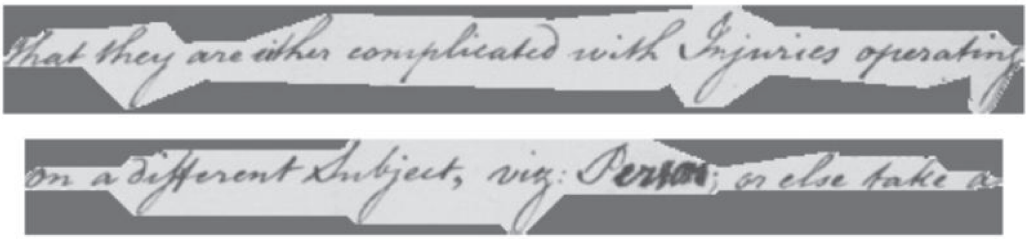


Figure 8. Sample lines such as they were provided for experimentation (Toselli and Vidal, 2015).

4.2. Experimental Setup

The first steps of a recognition system consist of preprocessing the input images and then extracting features. We applied preprocessing techniques to clean and enhance images, to correct skewed lines and slanted handwriting and normalized the size of images according to (Pastor et al., 2014, 2004a, 2006). A height of 64 pixels was defined for all text lines (keeping the aspect ratio) because we considered this number appropriate for representation and the application of the WT. Figure 9 shows some preprocessed text from text lines of the Bentham database.

For feature extraction, the Algorithm 1 described in Section "Wavelet-Based Descriptors for HTR" was applied. We defined a sliding window of width 32 and shifted 4 pixels

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 115 of 404

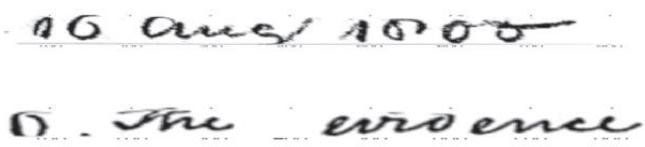


Figure 9. Preprocessed images from text lines of the Bentham database.

(we adjusted these values experimentally). The application of the CDF 9/7 to each window resulted in a feature vector of 512 coefficients in the case of the approximation subband at level 1 of the WT for the ***A1*** descriptor (see Subsection 3.2). The PCA transformation allowed reducing the feature vector to 24 components retaining the 89.24% of the data variance. In the case of ***A2***, 128 wavelet coefficients were obtained and reduced with PCA to 16, retaining 85.35% of the variance, while for ***A3***, 32 coefficients were obtained and reduced to 16 with PCA, retaining 92.87% of the variance. Reduction of descriptor size has a decisive impact on training time and on the processing of large databases and, sometimes allows an improvement in recognition percentages. For this reason, we considered a compromise between size and variance. Figure 10 depicts the process of feature extraction for descriptor ***A2*** on a text line image from Bentham database.

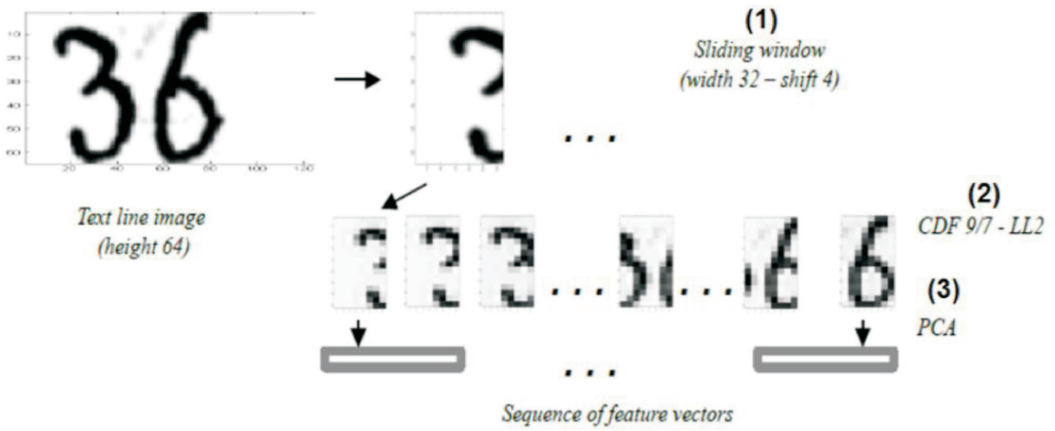


Figure 10. Feature extraction process for descriptor ***A2*** on a preprocessed text line image from Bentham database. (1) A frame is extracted with the sliding window; (2) WT is applied; (3) Size is reduced by PCA and the descriptor is obtained.

The recognition system used was the basic N-gram/HMM-GMM baseline system provided to the entrants of the ICFHR 2014 HTRtS competition (Causer and Wallace, 2012) implemented with the SRLIM (Stolcke, 2002) and HTK (Young et al., 2009) toolkits.

We chose a left-right HMM topology for all the characters. Each state has one transition to itself, and one to the next state (see example in Figure 1). Best results were obtained by defining the number of states of the HMM related to each character variable instead of setting the same number of states for all HMMs. For instance, it can be observed that some punctuation marks such as colon, semicolon, and parenthesis, are usually narrower

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 116 of 404

than other characters. Therefore, the number of states defined for the HMMs related to these alphabet elements is lower. As an example, we established a 3-state HMM for colon and semicolon, 4-state HMM for parenthesis, and 6-state HMM for the majority of letters. These values were set heuristically, and this is a topic to be better investigated in future work (Toselli and Vidal, 2015; Günter and Bunke, 2004). Finally, we built 88 HMMs (total number of alphabet elements) with 128 Gaussian densities per HMM state. These models were trained with the embedded Baum-Welch algorithm (Young et al., 2009), using the training and validation line images and their corresponding transcripts.

4.3. Results

For preliminary results, we trained the model through 5 iterations using a half of training and validation line images of the Bentham dataset used in ICHFR-2014 HTRtS contest (see Table 1), reducing the time of the training process considerably. The Word Error Rate (WER) is adopted to assess the HTR performance. WER is defined as the minimum number of words that need to be substituted, deleted or inserted to convert a recognized hypothesis into the corresponding reference transcription, divided by the total number of reference words (Pastor et al., 2014). Table 2 shows error percentages for each descriptor proposed in Section 3. It can be observed that values of WER are similar for the three descriptors. However, the size of feature vectors is lower for the A2 and A3 cases, speeding up training time.

Results were improved by using the entire training and validation sets for training and applying 20 iterations of BW algorithm. Also, training time was considerably increased. For descriptor ***A1***, a 26.19% of WER was obtained, while for ***A2*** and A3 WER values were 26.81% and 26.47% respectively. Error percentages were similar. However, training time for ***A2*** and ***A3*** was reduced by almost 50 % over the training time of ***A1***. This result is of considerable impact when the learning process lasts several days.

Table 2. N-gram/HMM-GMM HTR results on the ICFHR HTRtS contest Bentham dataset, using half of the training/validation line images with five iterations of BW algorithm (A), and the entire training set with 20 iterations of BW (B) for descriptors proposed and empirical settings outlined in Sec 4.2.

Descriptor	Dimension	WER (%)	
		(A)	(B)
A1	24	28.45	26.19
A2	16	28.49	26.81
A3	16	28.44	26.47

Table 3 compares our proposal with HTR results for Bentham dataset reported in the literature. The WT descriptors with a baseline recognition system outperform HTR percentages from published work (Bluche, 2015; Toselli and Vidal, 2015), obtaining the lower WER with A1. Additionally, data representation is improved through reducing the feature vector size by more than half in the case of ***A1*** and by more than 70 % in case of ***A2*** and ***A3***. Reduction of descriptor size has a decisive impact on training and processing times of large databases.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 117 of 404

Better results than ours were obtained by enhancing the N-gram/HMM-GMM system (tokenization and training algorithm) (Toselli and Vidal, 2015) using a moment-based descriptor of size 24. We believe that further improvements in data representation and recognition rates are possible to achieve by using WT descriptors with the enhanced N-gram/HMM-GMM system. We plan to apply these strategies for future work.

Table 3. N-gram/HMM-GMM HTR results on Bentham dataset using the sliding window framework.

N-gram/HMM-GMM approach	Descriptor	Dimension	WER (%)
ICFHR, HTRtS baseline Toselli and Vidal (2015)	Normalized,vertical gray levels and derivatives	60	35.30
HMM/GMM system Bluche Thesis (Bluche, 2015)	Handcrafted features	56	27.90
ICFHR HTRtS baseline Our proposal	**Wavelet-based A1** **Wavelet-based A2** **Wavelet-based A3**	24 16 16	**26.19** 26,81 26,47
ICFHR HTRtS enhanced tokenization Toselli and Vidal (2015)	Moment-based	24	23.90
ICFHR HTRtS enhanced tokenization + discriminative training, Toselli and Vidal (2015)	Moment-based	24	18.5

Conclusion

In this work descriptors for handwritten text recognition based on multiresolution features by the use of the CDF 9/7 Wavelet Transform and Principal Component Analysis are proposed. The approximation subbands from level 1 to 3 of the Wavelet Transform were considered for data representation because of the possibility of retaining the basic structure of the pattern achieving a strong reduction of the descriptor dimension. The feature vector is subject to PCA transformation to obtain a further reduction in size.

The recognition system is based on a segmentation-free approach which tightly integrate an optical character model and a language model that has yielded the best performance on standard benchmarks. The traditional N-gram/HMM-GMM model was adopted, implementing a baseline system trained with the embedded Baum-Welch algorithm. Experiments were performed on the challenging Bentham dataset used in the ICFHR 2014 HTRtS contest.

Our proposal outperformed HTR percentages reported in the literature for N-gram/HMM-GMM baseline systems. Additionally, data representation was improved as a result of reducing the feature vector size by more than 70 %. Reduction of descriptor size has a decisive impact on training and processing times of large databases. Better results than ours were obtained by enhancing the tokenization and training algorithm of the N-gram/HMM-GMM recognizer. In particular, the discriminative training strategy has

Cerence Operating Company, Ex. 2087
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 118 of 404

yielded good results. We plan to apply these strategies to our system for future work.

Acknowledgments

Work supported by the National Postdoctoral Program PNPD / CAPES of the government of Brazil during 2015. The authors would like to thank the CNPQ for supporting the development of this work through the research projects granted by "Edital Universal" (Process 444745/2014-9) and "Bolsa de Produtividade DT" (Process 311912/2015-0). The authors would also like to thank Alejandro Toselli, Moisés Pastor, and Enrique Vidal from the Pattern Recognition and Human Language Technology Research Center at Universidad Politécnica de Valencia for the advice provided.

References

Bazzi, I., Schwartz, R., and Makhoul, J. (1999). An omnifont open-vocabulary OCR system for english and arabic. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(6):495–504.

Bezerra, B. L. D., Zanchettin, C., and de Andrade, V. B. (2012). *A MDRNN-SVM Hybrid Model for Cursive Offline Handwriting Recognition*, pages 246–254. Springer Berlin Heidelberg, Berlin, Heidelberg.

Bluche, T. (2015). *Deep Neural Networks for Large Vocabulary Handwritten Text Recognition*. Theses, Université Paris Sud - Paris XI.

Bosch, V., Toselli, A. H., and Vidal, E. (2012). Statistical text line analysis in handwritten documents. In *2012 International Conference on Frontiers in Handwriting Recognition*. Institute of Electrical and Electronics Engineers (IEEE).

Causer, T. and Wallace, V. (2012). Building a volunteer community: results and findings from Transcribe Bentham. Digital Humanities Quarterly.

Chen, C.-M., Chen, C.-C., and Chen, C.-C. (2006). A comparison of texture features based on SVM and SOM.

Debnath, L. (2002). *Wavelet Transforms and Their Applications*. Springer Nature.

Dewangan, N. and Goswami., A. (2012). Image Denoising Using Wavelet Thresholding Methods. volume 2, pages 271–275.

El-Hajj, R., Likforman-Sulem, L., and Mokbel, C. (2005). Arabic handwriting recognition using baseline dependant features and hidden markov modeling. In *Proceedings of the Eighth International Conference on Document Analysis and Recognition*, ICDAR '05, pages 893–897, Washington, DC, USA. IEEE Computer Society.

Espana-Boquera, S., Castro-Bleda, M. J., Gorbe-Moya, J., and Zamora-Martinez, F. (2011). Improving offline handwritten text recognition with hybrid hmm/ann models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(4):767–779.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 119 of 404

Gouveia, F. M., Bezerra, B. L. D., Zanchettin, C., and Meneses, J. R. J. (2014). Handwriting recognition system for mobile accessibility to the visually impaired people. In *Systems, Man and Cybernetics SMC*, number 4 in 3, pages 3918–3981.

Günter, S. and Bunke, H. (2004). HMM-based handwritten word recognition: on the optimization of the number of states, training iterations and gaussian components. *Pattern Recognition*, 37(10):2069–2079.

Jelinek, F. (1998). *Statistical Methods for Speech Recognition*. MIT Press.

Kozielski, M., Forster, J., and Ney, H. (2012). Moment-based image normalization for handwritten text recognition. In *Proceedings of the 2012 International Conference on Frontiers in Handwriting Recognition*, ICFHR '12, pages 256–261, Washington, DC, USA. IEEE Computer Society.

Likforman-Sulem, L., Zahour, A., and Taconet, B. (2006). Text line segmentation of historical documents: a survey. volume 9, pages 123–138. Springer Nature.

Mallat, S. (1999). *A Wavelet Tour of Signal Processing*. Academic Press.

Marti, U.-V. and Bunke, H. (2002). Hidden markov models. chapter Using a Statistical Language Model to Improve the Performance of a HMM-based Cursive Handwriting Recognition Systems, pages 65–90. World Scientific Publishing Co., Inc., River Edge, NJ, USA.

Menasri, F., Likforman-Sulem, L., Mohamad, R. A.-H., Kermorvant, C., Bianne-Bernard, A.-L., and Mokbel, C. (2011). Dynamic and contextual information in hmm modeling for handwritten word recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:2066–2080.

Michal, Kozielski, Doetsch, P., and Ney, H. (2013). Improvements in rwth's system for off-line handwriting recognition. In *2013 12th International Conference on Document Analysis and Recognition*. Institute of Electrical and Electronics Engineers (IEEE).

Mohamad, R. A.-H., Likforman-Sulem, L., and Mokbel, C. (2009). Combining slanted-frame classifiers for improved hmm-based arabic handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(7):1165–1177.

Pastor, M., Sánchez, J., Toselli, A. H., and Vidal, E. (2014). Handwritten Text Recognition: Word-Graphs, Keyword Spotting and Computer Assisted Transcription.

Pastor, M., Toselli, A., and Vidal, E. (2004a). Projection profile based algorithm for slant removal. pages 183–190.

Pastor, M., Toselli, A., and Vidal, E. (2004b). Projection profile based algorithm for slant removal. pages 183–190.

Pastor, M., Toselli, A. H., and Vidal, E. (2006). Criteria for handwritten off-line text size normalization .

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 120 of 404

Patel, D. K., Som, T., Yadav, S. K., and Singh, M. K. (2012). Handwritten character recognition using multiresolution technique and euclidean distance metric. *Journal of Signal and Information Processing*, 03(02):208–214.

Romero, V., Toselli, A. H. and Vidal, E. (2012). Multimodal Interactive Handwritten Text Transcription. In *Series in Machine Perception and Artificial Intelligence (MPAI)*, World Scientific.

Sánchez, J. A., Bosch, V., Romero, V., Depuydt, K., and de Does, J. (2014). Handwritten text recognition for historical documents in the transcriptorium project. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, DATeCH '14, pages 111–117, New York, NY, USA. ACM.

Sánchez, J. A., Mühlberger, G., Gatos, B., Schofield, P., Depuydt, K., Davis, R. M., Vidal, E., and de Does, J. (2013). tranScriptorium. In *Proceedings of the 2013 ACM symposium on Document engineering - DocEng'13*. Association for Computing Machinery (ACM).

Sanchez, J. A., Toselli, A. H., Romero, V., and Vidal, E. (2015). ICDAR 2015 competition HTRtS: Handwritten text recognition on the tranScriptorium dataset. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. Institute of Electrical and Electronics Engineers (IEEE).

Seijas, L. M. and Segura, E. C. (2012). A wavelet-based descriptor for handwritten numeral classification. In *2012 International Conference on Frontiers in Handwriting Recognition*. Institute of Electrical and Electronics Engineers (IEEE).

Shao, Y. and Chang, C.-H. (2005). Wavelet transform to hybrid support vector machine and hidden markov model for speech recognition. In *2005 IEEE International Symposium on Circuits and Systems*. Institute of Electrical and Electronics Engineers (IEEE).

Skodras, A., Christopoulos, C., and Ebrahimi, T. (2001). JPEG2000: The upcoming still image compression standard. *Pattern Recognition Letters*, 22(12):1337–1345.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. in Proc. of ICSLP, Denver, USA.

Toselli, A. H., Juan, A., González, J., Salvador, I., Vidal, E., Casacuberta, F., Keyers, D., and Ney, H. (2004a). INTEGRATED HANDWRITING RECOGNITION AND INTERPRETATION USING FINITE-STATE MODELS. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(04):519–539.

Toselli, A. H. and Vidal, E. (2015). Handwritten text recognition results on the bentham collection with improved classical n-gram-hmm methods. In *Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing*, HIP '15, pages 15–22, New York, NY, USA. ACM.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

Trivedi, N., Kumar, V., Singh, S., Ahuja, S., and Chadha, R. (2011). Speech recognition by wavelet analysis. *International Journal of Computer Applications*, 15(8):27–32.

Young, S., Evermann, G., Gales, M., Hain, T., and Kershaw, D. (2009). The HTK Book: Hidden Markov Models Toolkit V3.4. Microsoft Corporation Cambridge Research Laboratory Ltd.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 122 of 404

In: Handwriting: Recognition, Development and Analysis ISBN: 978-1-53611-937-4
Editors: Byron L. D. Bezerra et al. © 2017 Nova Science Publishers, Inc.

Chapter 5

HOW TO DESIGN DEEP NEURAL NETWORKS FOR HANDWRITING RECOGNITION

***Théodore Bluche*[1,*], *Christopher Kermorvant*[2] *and Hermann Ney*[3]**
[1]A2iA SAS, Paris, France
[2]Teklia SAS, Paris, France
[3]RWTH Aachen University, Aachen, Germany

1. Introduction

We live in a digital world, where information is stored, processed, indexed and searched by computer systems, making its retrieval a cheap and quick task. Handwritten documents are no exception to the rule. The stakes of recognizing handwritten documents, and in particular handwritten texts, are manifold, ranging from automatic cheque or mail processing to archive digitalization and document understanding. The regions of the image containing handwritten text must be found, and converted into ASCII text, a process known as offline handwriting recognition.

This field has benefited from over sixty years of research. Starting with isolated characters and digits, the focus shifted to the recognition of words. The current strategy is to recognize lines of text directly, and use a language model to constrain the transcription, and help retrieve the correct sequence of words. One of the most popular approaches nowadays consists in scanning the image with a sliding window, from which features are extracted. The sequences of such observations are modeled with character Hidden Markov Models (HMMs). Word models are obtained by concatenation of character HMMs. The standard model of observations in HMMs is Gaussian Mixture Models (GMMs). In the nineties, the theory to replace Gaussian mixtures and other generative models by discriminative models, such as Neural Networks (NNs), was developed (Bourlard and Morgan, 1994). Discriminative models are interesting because of their ability to separate different HMM states, which improves the capacity of HMMs to differentiate the correct sequence of characters.

A drawback of HMMs is the local modeling, which fails to capture long-term dependencies in the input sequence, that are inherent to the considered signal. Recent improve-

*E-mail address: tb@a2ia.com.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

ments in Recurrent Neural Networks (RNNs), a kind of NN suited to sequence processing, significantly reduced the error rates. The Long Short-Term Memory units (LSTM, (Gers, 2001)), in particular, enable RNNs to learn arbitrarily long dependencies from the input sequence, by controlling the flow of information through the network. The current trend in handwriting recognition is to associate neural networks, especially LSTM-RNNs, with HMMs to transcribe text lines. NNs are used either to extract features for Gaussian mixture modeling (Kozielski et al., 2013a), or to predict HMM states and replace GMM optical models (Doetsch et al., 2014; Pham et al., 2014). On the other hand, in many machine learning applications, including speech recognition and computer vision, deep neural networks, consisting of several hidden layers produced a significant reduction of error rates.

Deep neural networks now get considerable interest in the machine learning community, and present many interesting aspects, *e.g.* their ability to learn internal representations of increasing complexity of their inputs, reducing the need of extracting relevant features from the image before the recognition. In the last few years, they have become a standard component of speech recognition models, which are close to those applied to handwriting recognition.

In this chapter, we focus on the hybrid NN/HMM framework, with optical models based on deep neural networks, for large vocabulary handwritten text line recognition. We concentrate on neural network optical models and propose a thorough study of their architecture and training procedure, but we also vary their inputs and outputs. We are interested in answering the following questions:

- *Is it still important to design handcrafted features when using deep neural networks, or are pixel values sufficient?*
- *Can deep neural networks give rise to big improvements over neural networks with one hidden layer for handwriting recognition?*
- *How (deep) Multi-Layer Perceptrons compare to the very popular Recurrent Neural Networks, which are now widespread in handwriting recognition and achieve state-of-the-art performance?*
- *What are the important characteristics of Recurrent Neural Networks, which make them so appropriate for handwriting recognition?*
- *What are the good training strategies for neural networks for handwriting recognition? Can the Connectionist Temporal Classification (CTC, (Graves et al., 2006)) paradigm be applied to other neural networks? What improvements can be observed with a discriminative criterion at the sequence level?*

The chapter will be divided as follows. In section "Experimental Setup", we describe the databases, neural networks, and the training and evaluation methods. In section "Hybrid Hidden Markov Model - Neural Network for Handwriting Recognition", we give an overview of the hybrid NN/HMM system. We present the components of the pipeline that will remain fixed throughout the rest of the chapter, namely the image pre-processing, the extracted features, the language models and the sliding window and HMM parameters. We also present baseline GMM/HMMs to validate those design choices. In sections "The

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 124 of 404

Impact of Inputs" through "The Impact of Outputs and Training Method", we present an experimental evaluation of many aspects of neural network optical models. We discuss the type of inputs in section "The Impact of Inputs", the network architectures in section "The Impact of Architecture" and we evaluate training methods and choice of outputs in section "The Impact of Outputs and Training Method". In section "Final Results", we select the best MLPs and RNNs, with features and pixel inputs, resulting from the conducted study. We evaluate the impact of the linguistic constraints (lexicon and language model), and the combination of these models. We compare the final results to previous publications and report state-of-the-art performance. The last section concludes this chapter by answering the proposed questions.

2. Experimental Setup

2.1. Databases

Table 1. Number of pages, lines, words and characters in each dataset

Database	Set	#Pages	#Lines	#Words	(unique)	#Characters	(unique)
Rimes *(French)*	*Train*	1,391	10,203	73,822	(8,061)	460,201	(97)
	Dev.	149	1,130	8,380		51,924	
	Eval.	100	778	5,639		35,286	
IAM *(English)*	*Train*	747	6,482	55,081	(7,843)	287,727	(79)
	Dev.	116	976	8,895		43,050	
	Eval.	336	2,915	25,920		128,531	
Bentham *(English)*	*Train*	350	9,198	76,707	(12,104)	419,764	(93)
	Dev.	50	1,415	11,580		64,070	
	Eval.	33	860	7,868		40,231	

The **Rimes** database (Augustin et al., 2006) consists of images of handwritten letters from simulated French mail. We followed the setup of the ICDAR 2011 competition. The available data are a training set of 1,500 paragraphs, manually extracted from the images, and an evaluation set of 100 paragraphs. We held out the last 149 paragraphs (approximately 10%) of the training set as a validation set and trained the systems on the remaining 1,391 paragraphs. Table 1 presents the number of words and characters in the different subsets. There are 460k characters distributed in more than 10k text lines, and 97 different symbols to be modeled (lowercase and capital letters, accentuated letters, digits and punctuation marks). The average character length, computed from the line widths, is 37.6 pixels at 300 DPI.

The **IAM** database (Marti and Bunke, 2002) consists of images of handwritten pages. They correspond to English texts extracted from the LOB corpus (Johansson, 1980), copied by different writers. The database is split into 747 images for training, 116 for validation, and 336 for evaluation. Note that this division is not the one presented in the official publication or on the website[1], but the one found in various publications (Bertolami and Bunke,

[1] http://www.iam.unibe.ch/fki/databases/iam-handwriting-database

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 125 of 404

2008; Graves et al., 2009; Kozielski et al., 2013b). We obtained the subdivision from H. Bunke, one of the creators of the database. Table 1 presents the number of words and characters in the different subsets. There are almost 290k characters distributed in more than 6k text lines, and 79 different symbols to be modeled (lowercase and capital letters, digits and punctuation marks). The average character length, computed from the line widths, is 39.1 pixels at 300 DPI.

The **Bentham** database contains images of personal notes of the British philosopher Jeremy Bentham, written by himself and his staff in English, around the 18th and 19th centuries. The data were prepared by University College, London, during the tranScriptorium project[2](Sánchez et al., 2013). We followed the setup of the HTRtS competition (Sánchez et al., 2014). The training set consists of 350 pages. The validation set comprises 50 images, and the test set 33 pages. Table 1 presents the number of words and characters in the different subsets. There are 420k characters distributed in almost 10k text lines, and 93 different symbols to be modeled (lowercase and capital letters, digits and punctuation marks). The average character length, computed from the line widths, is 32.7 pixels at 300 DPI.

2.2. Neural Networks

2.2.1. Multi-Layer Perceptrons

The perceptron (Rosenblatt, 1958) is a binary classifier, which goal is to take a *"yes/no"* decision. The output y can take two values, corresponding to a negative and positive decision, and can be formulated as

$$y = f(\mathbf{x}) = \sigma(b + w_1 x_1 + \ldots + w_n x_n) \tag{1}$$

where $\mathbf{x} = x_1, \ldots, x_n$ is an input feature vector, $b, w_1, \ldots, w_n$ are the free parameters (weights) and σ can be the *sigmoid* function:

$$\sigma(t) = \frac{1}{1 + e^{-t}} \tag{2}$$

Multi-Layer Perceptrons (MLPs) (Rumelhart et al., 1988)) are artificial neural networks, where the neurons are connected to each other. An MLP, as its name indicates, contains neurons organized in layers. Instead of the single perceptron, several neurons are connected to the same inputs $x_1, \ldots, x_n$, with a different set of weights. The outputs of all these neurons are inputs for a new layer of neurons.

The neurons of the last layer of the MLP are linear binary classifiers sharing the same input features. Thus an MLP with several outputs is a multi-class classifier. It was shown (Bourlard and Wellekens, 1989) that the outputs of the network can be interpreted as posterior probabilities. The *softmax* function (Bridle, 1990b) is often applied instead of the *sigmoid* function at the output layer. For n neurons with activations $a_1, \ldots, a_n$, the *softmax* function is defined as follows:

$$z_i = softmax(a_i) = \frac{e^{a_i}}{\sum_{k=1}^{n} e^{a_k}} \tag{3}$$

[2] http://transcriptorium.eu/

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 126 of 404

2.2.2. Recurrent Neural Networks

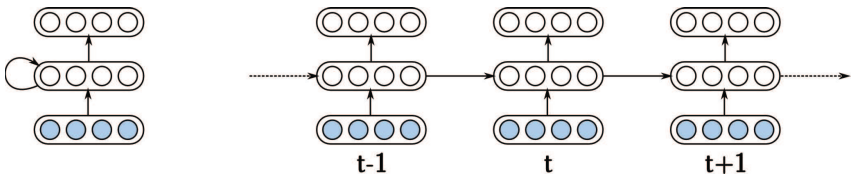


Figure 1. Recurrent Neural Networks, simple form.

Recurrent Neural Networks (RNNs) are networks with a notion of internal state, evolving through time, achieved by recurrent connections. In its simplest form, an RNN is an MLP with recurrent layers. A recurrent layer does not only receive inputs from the previous layers, but also from itself, as depicted on the left-hand side of Figure 1. The activations a_k^t of such a layer evolve through time with the following recurrence

$$a_k^t = \sum_{i=1}^{I} w_{ki}^{in} x_i^t + \sum_{h=1}^{H} w_{kh}^{rec} z_h^{t-1} \tag{4}$$

where x_is are the inputs and w_{ki}^{in} the corresponding weights, and z_h^{t-1} the layer's outputs at the previous timestep and w_{kh}^{rec} the corresponding weights.

Bidirectional RNNs (BRNNs, (Schuster and Paliwal, 1997)) process the sequence in both directions. In these networks, there are two recurrent layers: a forward layer, which takes inputs from the previous timestep, and a backward layer, connected to the next timestep. Both layers are connected to the same input and output layers.

2.2.3. Long Short-Term Memory Units

In RNNs, the vanishing gradient issue prevents the network to learn long time dependencies. (Hochreiter and Schmidhuber, 1997) proposed improved recurrent neurons called Long Short-Term Memory units. In LSTM, the flow of information is controlled by a gating system, scaling the input information, the output activation, and the contribution of the internal state of the unit at the previous timestep to the current state, based on the input and recurrent information and the cell internal state.

An LSTM cell is shown on Figure 2, and compared to a basic recurrent neuron. The cell input and all gates receive the activation of the lower layer and of the layer at the previous timestep.

With Long Short-Term Memory neurons in recurrent layers, Bidirectional and Multi-Dimensional RNNs achieve very good results in handwriting recognition, and constitute the state-of-the-art in that domain (Doetsch et al., 2014; Graves and Schmidhuber, 2008; Bluche et al., 2014; Moysset et al., 2014). In this chapter, we built Bidirectional LSTM-RNNs, with several recurrent layers. The two recurrent directions are merged with a feed-forward linear layer.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 127 of 404



Figure 2. Neurons for RNNs: (left) Simple Neuron (right) LSTM unit.

2.2.4. The Hybrid NN/HMM Scheme

In the hybrid approach (Bourlard and Morgan, 1994), GMMs are replaced by neural networks for the emission model of HMMs. The NN does not provide generative likelihoods $p(x_t|s)$, but discriminative state posteriors $p(s|x_t)$. We can use Bayes' rule:

$$p(x_t|s) = p(x_t)\frac{p(s|x_t)}{p(s)} \tag{5}$$

The joint probability defined by HMMs becomes:

$$\begin{aligned} p(\mathbf{W},\mathbf{x}) &= p(\mathbf{W})\sum_q\prod_t p(x_t|q_t)p(q_t|q_{t-1},\mathbf{W}) \\ &= \prod_t p(x_t)p(\mathbf{W})\sum_q\prod_t \frac{p(q_t|x_t)}{p(q_t)}p(q_t|q_{t-1},\mathbf{W}) \end{aligned}$$

H. Bourlard and his colleagues thoroughly studied the theoretical foundations of hybrid NN/HMM systems in (Bourlard and Wellekens, 1989; Bourlard and Morgan, 1994; Renals et al., 1994; Konig et al., 1996; Hennebert et al., 1997). In particular, they show in Konig et al. (1996) how a discriminant formulation of HMMs (Bourlard and Wellekens, 1989), able to compute $p(\mathbf{W}|\mathbf{x})$ leads to a particular MLP architecture predicting local conditional transition probabilities $p(q_t|q_{t-1},x_t)$, which allow to estimate global posterior probabilities.

2.3. Training Methods

2.3.1. Bootstrapping

One may train the neural network as a classifier, with a labeled dataset $\mathcal{S} = \{x^{(i)}, s^{(i)}\}$. In the hybrid NN/HMM approach, $x^{(i)}$s are frames and $s^{(i)}$s, HMM states. The targets may be obtained with uniform segmentation of observation sequences, or by alignment of the data using a trained system, *e.g.* GMM-HMM. One may re-align the observations with HMM states during the training procedure to refine the targets. The neural network is then plugged into the whole system, and its predictions provide scores for the decoding procedure. The

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 128 of 404

cost function is given by is

$$E_{NLL}(\mathcal{S}) = -\sum_{(x^{(i)},s^{(i)})\in\mathcal{S}} \log p(s^{(i)}|x^{(i)}) \tag{6}$$

2.3.2. Forward-Backward Training of Hybrid NN/HMM

The bootstrapping procedure presented above assumes a prior segmentation of the input data. However, an advantage of HMMs is the possibility to train and apply them to unsegmented data. In the Baum-Welch training, a forward-backward procedure is employed in the HMM of the true word sequence, in order to adjust the HMM parameters without making hard decisions about boundaries. Replacing the GMM likelihoods $p(x|s)$ by the scaled NN posteriors $\frac{p(s|x)}{p(s)}$ in the HMM formulation, and in forward and backward variables, one can apply the forward-backward algorithm to obtain state posterior probabilities in the HMM, including NN and transition probabilities. The cost function to be optimized is

$$E_{FwdBwd}(\mathcal{S}) = -\sum_{(\mathbf{x},\mathbf{z})\in\mathcal{S}} \log \sum_{\mathbf{s}\mapsto\mathbf{z}} \prod_t \frac{p(s_t|x_t)}{p(s_t)} p(s_t|s_{t-1}) \tag{7}$$

This training procedure, based on the forward-backward algorithm applied to HMMs or similar models can already be found in Alpha-nets (Bridle, 1990a). Bengio et al. (1992) and Haffner (1993) also propose a global training of the NN/HMM system using an MMI loss, computed with forward and backward variables. They report an improvement of results over the separate training of NN and HMMs. Senior and Robinson (1996) and Yan et al. (1997) first train a network with hard alignments, and then estimate the state posteriors with the forward-backward procedure to get a new, softer, estimate of targets, and use it for cross-entropy training of the NN. Konig et al. (1996) and Hennebert et al. (1997) focused on theoretical aspects, and explained the required assumptions for this training to achieve a global estimation of posterior probabilities.

2.3.3. Connectionist Temporal Classification

Connectionist Temporal Classification (CTC) was proposed by Graves et al. (2006), and corresponds to the task of labelling unsegmented data with neural networks. It is different from the previous methods, where there is one target at each timestep (or each frame in the sliding window approach). The basic idea of this framework is that the output of the neural network, when applied to an input sequence, is directly the sequence of symbols of interest, in our case the sequence of characters. The main advantages presented in Graves et al. (2006) are (i) that the training data does not need to be pre-segmented, *i.e.* we do not need one target for each frame to train the network, and (ii) that the output does not require any post-processing: it is already the sequence of characters, while usually neural networks predict posterior probabilities for HMM states, which should be decoded.

To make this possible, several artefacts are required. The input sequence has some length $T = |\mathbf{x}|$. Thus the length of the sequence of predictions (after the *softmax*) will also be T, while the length of the expected output sequence is generally smaller $|z| \leq T$. The simplest way to have the network predict characters directly is by removing duplicates in the output predictions, *e.g.* $AAAABBB \longrightarrow AB$. A problem arises when two successive

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 129 of 404

labels are the same, for example, if we want to predict AAB. This is one of the reasons why a *blank* symbol $\oslash$ is introduced in Graves et al. (2006), corresponding to observing no label. Therefore, in the CTC framework, the network has one output for each label in an alphabet L, plus one *blank* output, *i.e.* the output alphabet is $L' = L \cup \{\oslash\}$. A mapping $\mathcal{B} : {L'}^T \mapsto L^{\leq T}$ is defined, which removes duplicates, then blanks in the network prediction. For example: $\mathcal{B}(\oslash AA \oslash \oslash B) = \mathcal{B}(AAA \oslash BB) = AB$.

Provided that the network outputs for different timesteps are independent, given the input, the probability of a label sequence $\pi \in {L'}^T$ for a given $\mathbf{x}$ in terms of the RNN outputs is

$$p(\pi|\mathbf{x}) = \prod_t y^t_{\pi_t}(\mathbf{x}) \tag{8}$$

and the mapping $\mathcal{B}$ allows to calculate the posterior probability of a label (character) sequence $l \in L^{\leq T}$ by summing over all possible segmentations:

$$p(l|\mathbf{x}) = \sum_{\pi \in \mathcal{B}^{-1}(l)} p(\pi|\mathbf{x}) \tag{9}$$

Through the Equantion 9, we can train the network to maximize the probability of the correct labelling of the unsegmented training data $\mathcal{S} = \{(\mathbf{x}, z), z \in L^{\leq |\mathbf{x}|}\}$ by minimizing the following cost function

$$E_{CTC}(\mathcal{S}) = -\sum_{(\mathbf{x},\mathbf{z}) \in \mathcal{S}} \log p(z|\mathbf{x}) = -\sum_{(\mathbf{x},\mathbf{z}) \in \mathcal{S}} \log \sum_{\mathbf{s} \mapsto \mathbf{z}} \prod_t p(s_t|\mathbf{x}) \tag{10}$$

The computation of $p(z|\mathbf{x})$ implies a sum over all paths in $\mathcal{B}^{-1}(z)$, each of which of length $T = |\mathbf{x}|$, which is expensive. Graves et al. (2006) propose to use the forward-backward algorithm in a graph representing all possible labelling alternatives.

2.3.4. Sequence-Discriminative

Sequence-discriminative training optimizes criteria to increase the likelihood of the correct word sequence, while decreasing the likelihood of other sequences. This kind of training is similar to the discriminative training of GMM-HMMs with the Maximum Mutual Information (MMI) or the Minimum Phone Error (MPE) criteria.

The MMI criterion is defined as follows:

$$\mathcal{E}_{MMI}(\mathcal{S}) = \sum_{(\mathbf{x},\mathbf{W_r}) \in \mathcal{S}} \log \frac{p(\mathbf{W}_r|\mathbf{x})}{\sum_{\mathbf{W}} p(\mathbf{W}|\mathbf{x})} \tag{11}$$

The Minimum Phone Error (MPE, (Povey, 2004)) class of criteria has the following formulation:

$$\mathcal{E}_{MBR}(\mathcal{S}) = \sum_{(\mathbf{x},\mathbf{W_r}) \in \mathcal{S}} \frac{\sum_{\mathbf{W}} p(\mathbf{W}|\mathbf{x}) \mathbf{A}(\mathbf{W}, \mathbf{W_r})}{\sum_{\mathbf{W}'} p(\mathbf{W}'|\mathbf{x})} \tag{12}$$

where $A(\mathbf{W}_1, \mathbf{W}_2)$ is a measure of accuracy between $\mathbf{W}_1$ and $\mathbf{W}_2$. It is the number of correct characters for MPE or the number of correct HMM states in the recognized sequence (compared to the forced alignments) for state-level Minimum Bayes Risk (sMBR, (Kingsbury, 2009)). These criteria involve a summation over all possible word sequences, which

Cerence Exhibit 2014
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 130 of 404

is difficult to compute in practice. Instead, recognition lattices are extracted with the optical and language models, and only word sequences in these lattices are considered in the cost function.

Sequence-discriminative training is popular in hybrid NN/HMM in speech recognition. As already mentioned earlier, (Bengio et al., 1992; Haffner, 1993) applied the MMI criterion to the global training of a NN/HMM system. In the past few years, these training methods arouse much interest with the advent of deep neural networks (Kingsbury, 2009; Sainath et al., 2013; Veselý et al., 2013; Su et al., 2013). Usually, a neural network is first trained with a framewise criterion. Then, lattices are generated, and the network is further trained with MMI, MPE, or sMBR. Regenerating the lattices during training may be helpful, but the gains are limited beyond the first epoch (Veselý et al., 2013). In speech recognition, experiments with sequence-discriminative training yielded relative WER improvements in the range of 5-15%.

2.4. Evaluation

We carried out a thorough evaluation of different aspects of neural networks for handwriting recognition, with a particular focus on deep neural networks. We tried, as much as possible, to compare shallow and deep networks on the one hand, and feature and pixel inputs on the other hand. We evaluated several aspects and design choices for neural networks, including inputs, output space, training method, depth and architecture. All our experiments were conducted on the three databases (Rimes, IAM and Bentham).

Unless stated otherwise (in section "The Impact of Outputs and Training Method"), the MLPs are trained with the bootstrapping method, to classify each frame into one of the HMM state. The targets states are obtained with forced alignment of the training set with the baseline GMMs. The minimized cost is the one defined in Equation 6. The performance of the MLPs alone is evaluated with the ***Frame Error Rate (FER%)***, defined as the ratio of incorrectly classified frames over the total number of frames.

The RNNs are trained with the CTC method to directly predict character sequences, by minimizing the cost function defined in Equation 10. The performance of the RNNs alone is evaluated with the ***Character Error Rate (RNN-CER%)***, defined as the edit distance between the ground-truth and predicted character sequences, normalized by the number of characters in the ground-truth.

Keeping in mind that the networks will be used in a complete pipeline, we also measured the ***Character (CER%)*** and ***Word Error Rates (WER%)***, defined similarly as the RNN-CER%, after the integration of the language models.

3. Hybrid Hidden Markov Model - Neural Network System for Handwriting Recognition

In this chapter, we focus on the optical model of HMMs for handwriting recognition. More particularly, we study two kinds of deep neural networks in the hybrid NN/HMM framework: Multi-Layer Perceptrons and Recurrent Neural Networks. The inputs of our systems are sequences of feature vectors extracted from preprocessed line images. The outputs are posterior probabilities of HMM states.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 131 of 404

We explore different aspects of the neural networks: their structure, their parameters, and their training procedure. We present results of the whole recognition systems. Most of the components of these systems, excluding the neural networks, are kept fixed throughout the experiments, unless stated otherwise. This section is dedicated to the presentation of the fixed components: text line image preprocessing, feature extraction, modeling with Hidden Markov Models (HMMs), and language models.

Figure 3 shows an overview of the recognition system and of its components. The ones with thick, dashed lines are the fixed ones, presented in this section. In this section, we will also present a baseline optical model – a Gaussian Mixture Model.

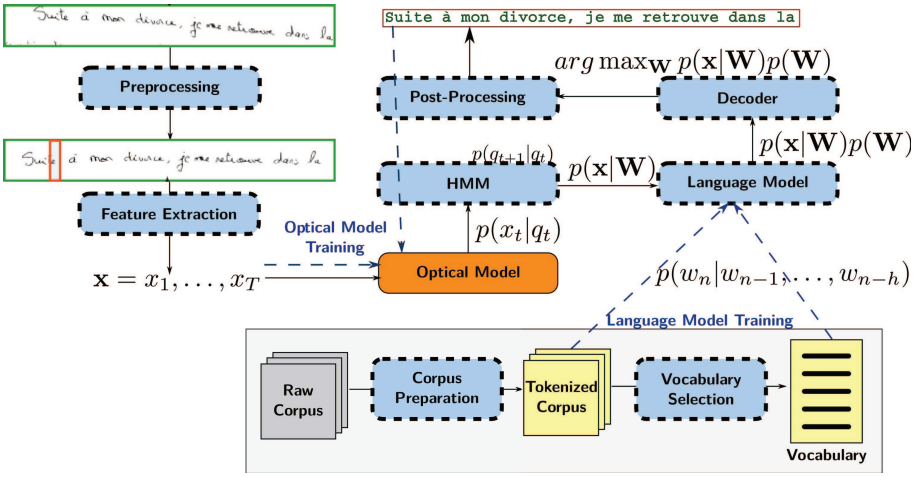


Figure 3. Overview of the recognition system.

3.1. Preprocessing and Feature Extraction

In this section, we present the image preprocessing applied and the features extracted. We experimented several options. Quick experiments consisted in training GMM/HMM systems with 10% of the training set, for only a few iterations and Gaussians per state, and recording the word error rate (WER) on 10% of the validation set, with a very small closed vocabulary and a unigram language model. We used the handcrafted features described in the second part of this section, extracted with a sliding window. We present results on IAM, where we tried most of the configurations. We also tuned Rimes and Bentham systems, starting with setups that were good for IAM. High WERs are caused by the limited amount of data (image and vocabulary) and order of language model, but the selected methods produced reasonable GMM/HMM baselines in the end.

3.1.1. Image Preprocessing

First the potential skew in the image is corrected with the algorithm of Bloomberg et al. (1995). We applied the slant correction method presented in Buse et al. (1997). For contrast enhancement, we tried adaptive thresholding (Otsu, 1979) and the interpolation method

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 132 of 404

Table 2. Selection of contrast enhancement method (%WER).

	Window size:	6px	9px
Method	None	54.2%	58.0%
	Adaptive	57.2%	58.5%
	Interpolation	**53.1%**	57.2%

Table 3. Selection of height normalization method (%WER).

	Window size:	6px	9px
Method	None	56.9%	59.6%
	Fixed (72px)	54.2%	58.7%
	Region (22px, 33px, 17px)	58.7%	63.8%
	Region (24px, 24px, 24px)	**53.1%**	57.2%

from Roeder (2009). The results reported in Table 2 show that the latter method is better, for two different sliding window size.

We also tried to normalize the height of images, either to a fixed value of 72px or with region-dependent methods (Toselli et al., 2004; Pesch et al., 2012), with fixed height for each region (ascenders, core, descenders – 22, 33 and 17px, or 24px for each). The regions are found after deskew and deslant with the algorithm of Vinciarelli and Luettin (2001). We selected the normalization of each region to 24px, based on the results (WER) of Table 3.

3.2. Feature Extraction with Sliding Windows

We used two kinds of features: handcrafted features, and raw pixel intensities. A sliding window is scanned through the line image to extract features.

3.2.1. Handcrafted Features

The handcrafted features are geometrical and statistical features extracted from the window. They were proposed by Bianne-Bernard (2011); Bianne et al. (2011), and derived from the work of El-Hajj et al. (2005). They gave good performance on several public databases (Menasri et al., 2012; Bianne et al., 2011).

The text baseline and core region are computed with the algorithm of Vinciarelli and Luettin (2001), and the following values are calculated:

- 3 pixel density measures: in the whole window, and in the regions above and below the baseline,
- pixel densities in each column of pixels (w_f values, where w_f is the width of the sliding window),
- 2 measures of the center of gravity: relative vertical positions with respect to the baseline and to the center of gravity in the previous window,

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 133 of 404

- 12 measures (normalized counts) of local pixel configurations: six configurations, computed from the whole window and from the core region,
- Histogram of Oriented Gradients (HOG) in 8 directions.

All these features form a $(25 + w_f)$-dimensional vector, to which deltas are appended, resulting in feature vectors of dimension 56 ($w_f = 3$px). The parameters of the sliding window (width and shift) for the handcrafted features have been tuned using the same method as for preprocessing. The best parameters were a shift and width of 3px each (no overlap between windows).

3.2.2. Pixel Values

The "pixel features" are extracted with a sliding window. The width of the window was optimized for deep neural networks: 45px for Rimes and IAM, 57px for Bentham. We also tried variations of these values in specific experiments. In order to extract the same number of frames for both kinds of features, the shift was fixed to be the same as for handcrafted features. To limit the number of features, each frame is downscaled from a height of 72px to a height of 32px. The aspect ratio is kept constant (20x32px for Rimes and IAM, 25x32px for Bentham).

3.3. Language Models

Each database has specificities. For example, Rimes contains many reference codes and acronyms. Hyphenation appears a lot in Bentham database. We applied different tokenizations to take them into account and limit the size of the vocabularies, such as separating the punctuation and codes.

For IAM, the Language Model (LM) is trained on the LOB (Johansson, 1980), Wellington (Janet Holmes and Johnson, 1998) and Brown (W. N. Francis, 1979) corpora, with a vocabulary made of the 50,000 most frequent words. For Rimes, the LM was trained on the tokenized training set paragraph annotations, with a vocabulary including all tokens (5,000). For Bentham, we extracted a vocabulary of 7,318 words from this corpus. In order to recognize hyphenated words, we added hyphenated versions in the vocabulary. For all words with more than ten occurrences, we generated all possible (beginning, end) pairs using Pyphen[3], and added the three possible hyphenation symbols at the end (*resp.* beginning) of words beginnings (*resp.* endings), and included them in the vocabulary. It increased the size to 32,692 words, but decreased the Out-Of-Vocabulary (OOV) rate on the validation set from 7.1 to 5.6%.

We trained language models for each database with the SRILM toolkit (Stolcke, 2002). For **IAM**, we used a 3gram language model trained on the tokenized LOB, Brown and Wellington corpora, with modified Kneser-Ney smoothing. The resulting model has a perplexity of 298 and an OOV rate of 4.3% on the validation set (respectively 329 and 3.7% on the evaluation set). For **Rimes**, we built a 4gram LM with Kneser-Ney discounting (Kneser and Ney, 1995). The language model has a perplexity of 18 and OOV rate of 2.9% on the validation set (respectively 18 and 2.6% on the evaluation set). For **Bentham**, we estimated

[3] http://pyphen.org

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 134 of 404

the LMs with the *n*gram counts from the corpus. The hyphenated word chunks are added to the unigrams with count 1. We generated 4grams with Kneser-Ney discounting (Kneser and Ney, 1995). Table 4 presents the perplexities of different *n*grams. They are better without hyphenation, but we found that the hyphenated version gave better recognition results.

Table 4. Perplexities of Bentham LMs with different *n*gram orders and hyphenation, on the validation set.

Hyphenation	Size	OOV%	1gram	2gram	3gram	4gram
No	7,318	7.1	348.7	129.4	**101.7**	**96.7**
Yes	32,692	5.6	656.1	137.6	108.4	**103.1**

We evaluated the system by comparing the recognition outputs to the ground-truth transcriptions. We decoded with the tools implemented in the Kaldi speech recognition toolkit (Povey et al., 2011), which consists of a beam search in a Finite-State Transducer (FST), with a token passing algorithm. The FST is the composition of the representation of each component (HMM, vocabulary and Language Model) as FSTs (H, L, G). The HMM and vocabulary conversion into FST is straightforward. The LM generated by SRILM is transformed by Kaldi. The final graph computation not only involves composition, but also other FST operations. The method proposed in Kaldi is a variation of the technique explained in Mohri et al. (2002):

$$F = min(rm(det(H \circ min(det(L \circ G))))) \tag{13}$$

where $\circ$ denotes FST composition, and *min*, *det* and *rm* are respectively FST minimization, determination, and removal of some ε-transitions and potential disambiguation symbols. Refer to Mohri et al. (2002); Povey et al. (2011) for more details concerning the FST creation.

3.4. Baseline GMM/HMM System

We chose a left-right HMM topology for all the characters. Each state has one transition to itself, and one to the next state. The whitespace is modeled by a 2-state HMM. All other character HMMs have the same number of states, tuned along with the sliding window topology. Each state if associated with its own emission probability. We built 96 character HMMs with 5 states for Rimes, 78 character HMMs with 6 states for IAM, and 92 character HMMs with 6 states for Bentham.

The goals of these GMM/HMMs are: (*i*) to check that we chose a good preprocessing, feature extraction, and HMM topology, (*ii*) to serve as a baseline to compare hybrid models with, and (*iii*) to produce forced alignments to build training sets for neural networks. The results are presented on Tables 5 (Rimes) and 6 (IAM). We compare them to the best published results with pure GMM/HMMs and with other systems. On Rimes (Table 5), some publications do not report results on the development set, while others, such as Kozielski et al. (2014), were directly tuned on the evaluation set. Yet our GMM/HMM system achieves WER and CER competitive with the GMM/HMM of Kozielski et al. (2014). The results are also reasonable on IAM (Table 6). The first line of the comparison (Kozielski

Cerence Operating Company, Exhibit 1017
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 135 of 404

Table 5. Results on Rimes database

	Dev.		Eval.	
	WER	**CER**	**WER**	**CER**
Our GMM/HMM	17.2	5.9	15.8	6.0
GMM/HMM systems				
(Kozielski et al., 2014)	-	-	15.7	5.5
(Grosicki and El-Abed, 2011)	-	-	31.2	18.0
Other systems				
(Pham et al., 2014)	-	-	12.3	3.3
(Doetsch et al., 2014)	-	-	12.9	4.3
(Messina and Kermorvant, 2014)	-	-	13.3	-

et al., 2013b) uses an open-vocabulary approach, able to recognize any word (no OOV). On

Table 6. Results on IAM database

	Dev.		Eval.	
	WER	**CER**	**WER**	**CER**
Our GMM/HMM	15.2	6.3	19.6	9.0
GMM/HMM systems				
(Kozielski et al., 2013b)	12.4	5.1	17.3	8.2
(Kozielski et al., 2014)	12.6	4.7	-	-
(Kozielski et al., 2013b)	18.7	8.2	22.2	11.1
(Toselli et al., 2010)	-	-	25.8	-
(Bertolami and Bunke, 2008)	26.8	-	32.8	-
Other systems				
(Doetsch et al., 2014)	8.4	2.5	12.2	4.7
(Kozielski et al., 2013a)	9.5	2.7	13.3	5.1
(Pham et al., 2014)	11.2	3.7	13.6	5.1

Bentham database, we obtained a WER of 27.9% and a CER of 14.5%

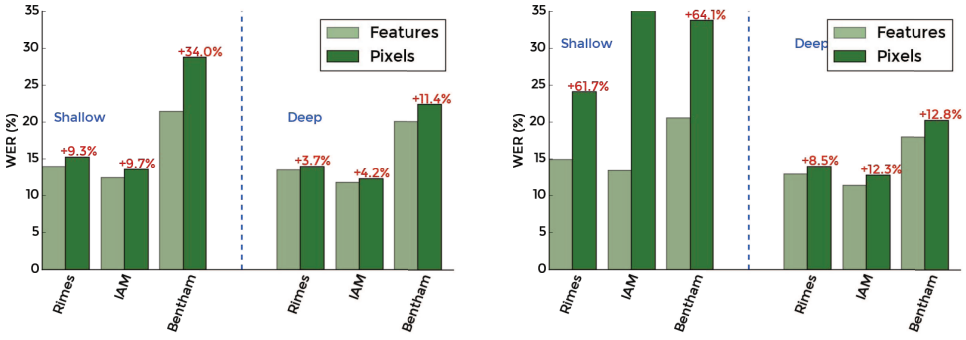
4. The Impact of Inputs

In this first section of experiments, we focus on the impact of the inputs given to the neural network on the performance. More specifically, we compare handcrafted features to pixels values, and we evaluate the importance of providing contextual information to the network.

4.1. Types

In the introduction, we presented two kinds of inputs for the neural networks: handcrafted features and pixel values. In many pattern recognition problems, the advent of deep neural networks allowed to replace handcrafted features by the raw signal. The relevant features are learnt by the recognition system. Using raw inputs has some advantages, such as relieving the architect of the system from implementing feature extraction methods.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 136 of 404



(a) MLPs (b) RNNs

Figure 4. Comparison of pixels and handcrafted features as inputs to shallow and deep MLPs and RNNs.

On Figure 4, we plot the WER% of complete hybrid systems, in which the optical model is an MLP or an RNN, with one or several hidden layers. We compare the performance with handcrafted features and pixel values. Although the raw pixel values always seem a little worse than the handcrafted features, we observe a big reduction of the performance gap when using deep neural networks. This is especially striking for recurrent neural networks, which give a high WER% with pixels and only one hidden layer. We will see later in this chapter that when using better training methods, such as the sequence-discriminative training and dropout, the gap almost disappears.

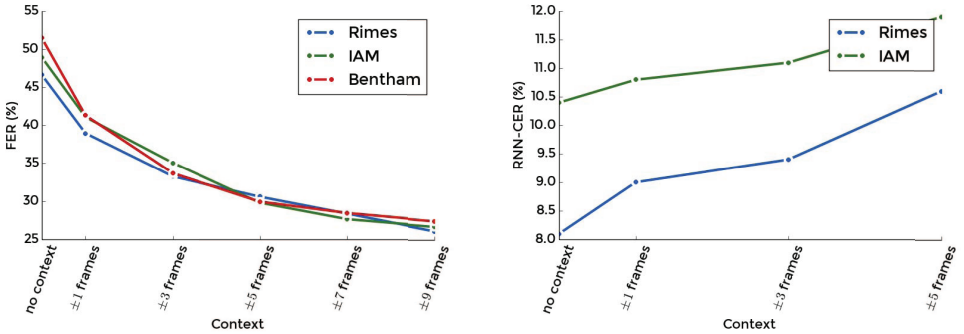
4.2. Context

4.2.1. Context through Frame Concatenation

The inputs of the neural networks are sequences of feature vectors, extracted with a sliding window. This window is usually relatively small, ofter smaller than a character. For handcrafted features, increasing the size of the window would result in a probable loss of information, as more pixels will be summarized in the same number of features. A common alternative approach to providing more context to the neural networks is to concatenate successive frames.

We report the results of that experiment in Figure 5. On the lefthand side, for MLPs, we observe that the Frame Error Rate (FER) decreases when more context is provided (from around 50% without context to less than 30% with a lot of context). The improvements are not as big in the complete system because the HMM and language model help to correct many mistakes made by the MLP. Yet we observed up to 20% relative WER% improvement by choosing the right amount of context. On the other hand, we notice a performance drop when explicitly adding context in the inputs of RNNs (Figure 5b). Although surprising, it should be noted that concatenating frames increases the dimension of the input space. While MLPs classify frames independently and need more context to compensate small frames, RNNs process the whole sequence, and can learn to propagate context from adjacent frames, as we will see in the next section.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 137 of 404



(a) MLPs (b) RNNs

Figure 5. Benefits of concatenating successive frames to provide more context to neural networks.

4.2.2. Context through the Recurrent Connections

In this section, we try to see how RNNs incorporate the context to predict characters through the recurrent connections. Similarly as in Graves et al. (2013), we observed the sensitivity of the output prediction at a given time to the input sequence.

To do so, we computed the derivative of the RNN output y_τ at time $t = \tau$ with respect to the input sequence $\mathbf{x}$. We plotted the sensitivity map $S = (S_{t,d})_{1 \leq t \leq |\mathbf{x}|, 1 \leq \mathbf{d} \leq \mathbf{D}}$, where D is the dimension of the feature vector, and:

$$S_{t,d} = \left| \frac{\partial y_\tau}{\partial x_{t,d}} \right| \tag{14}$$

For pixel inputs, the sliding windows are overlapping, and the dimension of the feature vectors is $D = wh$, where w is the width of the window, and h its height. Therefore, we can reshape the feature vector to get the shape of the frame, and a sensitivity map with the same shape as the image. In overlapping regions, the magnitude of the derivative for consecutive windows are summed:

$$S_{i,j} = \sum_{k=-w/2\delta}^{w/2\delta} \left| \frac{\partial y_\tau}{\partial x_{i+\delta k, i+w(j-1)-\delta k}} \right| \tag{15}$$

where $\delta = 3px$ is the step size of the sliding window. This way, we can see the sensitivity in the image space.

On Figure 6, we display the results for BLSTM-RNNs with 7 hidden layers of 200 units. On each plot, we show on top the preprocessed image, the position τ and the RNN prediction y_τ, as well as the sliding window at this position to put the sensitivity map in the perspective of the area covered by the window at $t = \tau$.

The step size δ of all sliding windows is 3px, *i.e.* the size of the sliding window for features displayed on the top plots. We observe that the input sensitivity goes beyond ± 5 frames, as well as beyond the character boundaries in some cases, as if the whole word could help to disambiguate the characters. It is also an indication that RNNs actually use their ability to model arbitrarily long dependency, an ability that MLPs lack.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 138 of 404

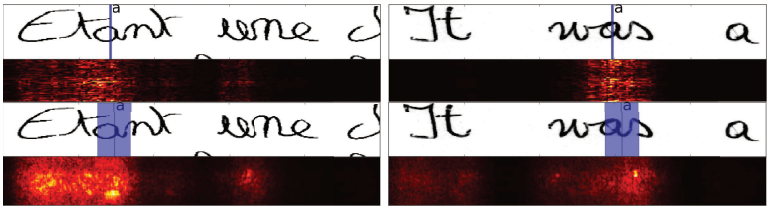


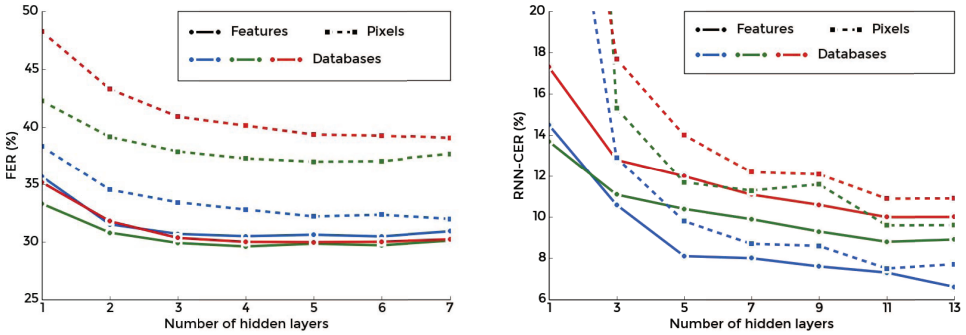
Figure 6. Context used trough recurrent connections by LSTM-RNNs to predict character "a" (sensitivity heatmaps, top: features, bottom: pixels).

5. The Impact of Architecture

In this section, we focus more closely on the network itself, and in particular its architecture. There are several design choices to make, including the number and types of hidden layers. Since this chapter is about deep neural networks, we will first measure the influence of depth for the two kinds of neural networks, and see that deeper networks tend to perform better. Then, we will observe that the impact of recurrent layers in the proposed RNNs is bigger in upper layers.

5.1. Depth

The first experiment consists of adding hidden layers and measure the effect on the performance of the neural network, outside of the complete pipeline. We measured the FER% for MLPs and the RNN-CER% for RNNs (trained with CTC). The MLPs have 1,024 neurons in each hidden layer. The RNNs have 200 units in each layer. For RNNs, we actually add a BLSTM layer, i.e. 200 units in each scanning direction, plus a linear layer with 200 units too to merge the information coming from both directions.



(a) MLPs (b) RNNs

Figure 7. Effect of increasing the number of hidden layers in the performance of MLPs and RNNs alone.

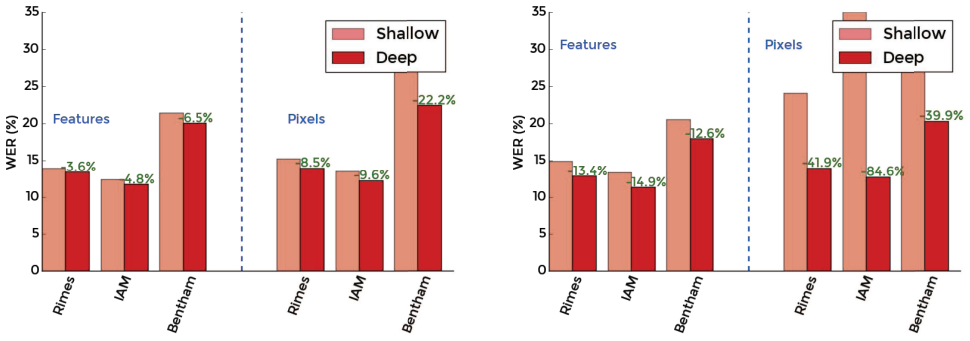
The results are displayed in Figure 7, for MLPs and RNNs, on all three databases, with pixel and handcrafted features as inputs. For MLPs, we notice relative FER improvements up to 20% going from one to several hidden layers. The biggest improvement is observed

Cerence Operating Company, Ex. 2004
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 139 of 404

from one to two hidden layers, but we still get better results with more layers. Overall, four or five hidden layers look like a good choice to get optimal FER: the improvements beyond that number are relatively small.

For RNNs, we observe that almost every time we add layers, the performance of the RNN is increased. For handcrafted features, adding a second LSTM layer and a feed-forward one brings around a relative 20-25% CER improvement. Adding a third one yields another 6-12% relative improvement. For pixels, one hidden layer only is not sufficient, and adding another LSTM layer divides the error rates by more than two. A third LSTM layer is also significantly better, by another 20-25% relative CER improvement.

In Figure 8, we show the WER% results when shallow 1-hidden layer and deep networks are included in the complete pipeline. The obtained improvement are not as impressive as those of the networks evaluated alone, but remain significant, especially with pixels as inputs. This is particularly striking for RNNs, which yield high error rates when there is only one hidden layer, but which achieve similar results as the feature-based RNNs with several hidden layers.



(a) MLPs (b) RNNs

Figure 8. Comparison of recognition results (WER%) of the full pipeline with shallow (one hidden layer) and deep neural networks.

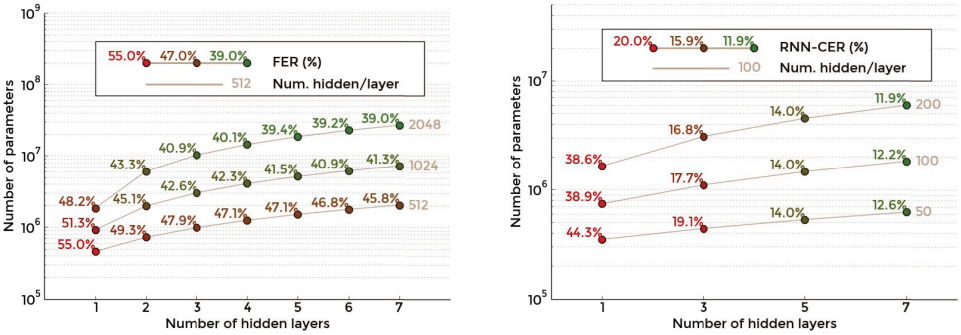
It should be noted that adding hidden layers increases the total number of free parameters, hence the global capacity of the network. We may also control the number of free parameters by varying the number of neurons in each layer. In Figure 9, we show the number of parameters and error rates when changing the depth on the one hand, and the number of neurons on the other hand. We see that for a fixed number of parameters, deeper networks tend to perform better.

From these experiments, we can conclude that depth, not only the number of free parameters plays an important role in the reduction of error rates.

5.2. Recurrence

In this second set of experiments, we measure the impact of recurrence in the neural networks. The results are summarized in Figure 10. As one can notice, similar error rates are achieved by the two kinds of optical models, and of inputs, making a definite conclusion

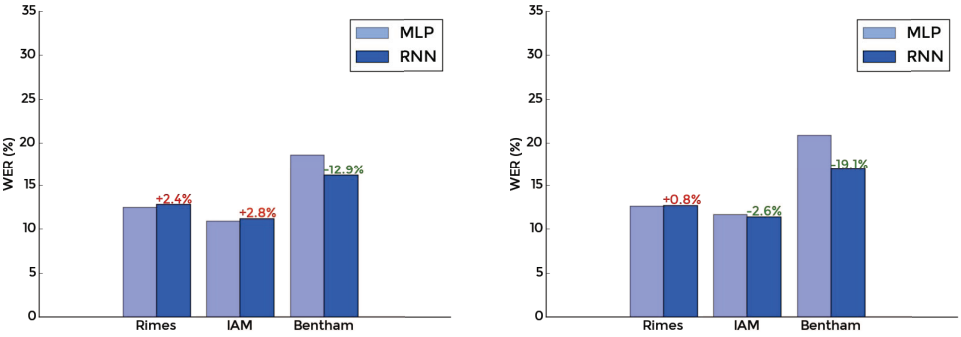
Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 140 of 404



(a) MLPs (b) RNNs

Figure 9. Comparison of performance of neural networks when the number of free parameters is adjusted by varying the depth and the number of neurons in hidden layers.

hard to draw about what are the best choices. Yet, since pixel values yield similar performance as handcrafted features, the need to design and implement features vanishes, and one may simply use the pixels directly. Moreover, although RNNs are found in all the best published systems for handwritten text line recognition, they are not the only option, and MLPs should not be neglected.



(a) Handcrafted (b) Pixels

Figure 10. Comparison of MLPs and RNNs, for both kinds of inputs, in the complete pipeline.

Next, we replace LSTM layers by feed-forward ones in BLSTM-RNNs made of five hidden layers (three LSTM and two feed-forward layers). To keep the number of parameters approximately constant, we replace the original LSTM layer with 100 units for each direction by a feed-forward layer with 900 units. We trained all eight possible combinations on Rimes and IAM, with pixel and feature inputs. We refer to the different architectures by a triplet indicating the type of layer at the original LSTM positions (bottom, middle, top). **RRR** represents the original RNNs, while FFF corresponds to a purely feed-forward MLP.

The results are exposed in Table 7. If we look at the different positions of one LSTM layer (**R**FF, F**R**F and FF**R**), in almost all cases, the higher it is in the network, the better the

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 141 of 404

Table 7. Effect of recurrence on the character error rate of the RNN alone (RNN-CER%)

	Features		Pixels	
	Rimes	IAM	Rimes	IAM
FFF	44.0	39.6	38.0	32.8
RFF	13.2	13.7	62.2	61.3
F**R**F	12.3	13.7	20.6	19.2
FF**R**	13.0	12.5	17.5	**17.5**
RRF	11.6	23.1	20.8	20.3
RF**R**	11.6	11.8	23.0	19.6
F**RR**	11.6	12.0	**15.3**	**17.5**
RRR	**9.7**	**11.4**	16.7	18.9

results are. This is especially visible with pixel inputs.

Adding LSTM layers seems generally helpful, although it is not always the case. For example, adding LSTM in the first hidden layer degrades a lot the performance with pixel inputs. It might be due to the fact that for pixels, the lower layers extract elementary features from the image. On the other hand, recurrence seems important in the CTC framework, as shown in the next section. Therefore, it is probably too difficult for this first layer to learn both the low-level features required to interpret the image and the dependency that is necessary for the convergence of the CTC.

6. The Impact of Outputs and Training Method

6.1. Dropout

The dropout technique (Hinton et al., 2012) is a regularization method for neural networks. It consists in randomly setting some activations from a given hidden layer to zero during training. Pham et al. (2014) and Zaremba et al. (2014) recently proposed a way to use dropout in LSTM networks. Pham et al. (2014) carried out experiments on handwritten word and line recognition with MDLSTM-RNNs, and reported relative WER improvements between 2 and 15% for line recognition with language models, and also observed effects on the classification weights similar to those of L_2 regularization.

The authors proposed to apply dropout with $p = 0.5$ in feed-forward connections following the LSTM layers, and show that in an architecture of three levels of LSTM layers, it is generally better to use dropout after every LSTM layer, rather than after the last one or two only. Here we propose to explore the dropout technique within our deep BLSTM-RNN architecture. We experimented dropout at different positions, depicted on Figure 11, that is either before the LSTM layer, after it, or in the recurrent connections. Moreover, we studied the effect of dropout in different layers in isolation.

In Figure 12, we report the relative RNN-CER% improvement brought by dropout at different places, compared to the network without dropout. Looking at individual configurations (top-left plot), it is hard to draw a general conclusion about where is the best place to apply dropout. Yet, besides the fact that dropout almost always helps, we can draw several

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 142 of 404

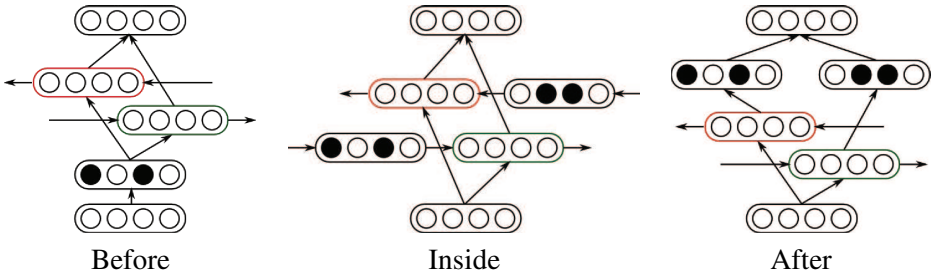


Figure 11. Dropout position in LSTM-RNNs.

conclusions. When dropout is only applied at one position (top plots of Figure 12):

- it is generally better in lower layers of the RNNs, rather than in the top LSTMs, except when it is after the LSTM.
- it is almost always better before the LSTM layer than inside or after it, and better after than inside, except for the bottom layer.

When it is applied in all layers (bottom, middle and top; bottom plot of Figure 12):

- among all relative positions to the LSTM, when dropout is applied to every LSTM, placing it after was the worst choice in all six configurations
- before LSTMs seems to be the best choice for Rimes and Bentham, and inside LSTMs is better for IAM.

In the complete pipeline, with language models, we studied the results with dropout at different positions relative to the LSTM layers (Figure 13). We observe that for Rimes, the best results are achieved with dropout after LSTMs, despite the superior performance of dropout before for the RNN alone. For IAM, dropout inside LSTMs is only slightly better for features. With pixel inputs, placing dropout before LSTMs seems to be a good choice. The main difference between the RNN alone and the complete system is that the former only considers the best character hypothesis at each timestep, whereas the latter potentially considers all predictions in the search of the best transcription with lexical constraints.

Therefore, applying dropout after the LSTM in the top layer(s) might be beneficial for the beam search in the decoding with complete systems. Indeed, dropout after the last LSTMs forces the classification to rely on more units. Conversely, a given LSTM unit will contribute to the prediction of more labels. On the other hand, the values of neighboring pixels are highly correlated. If the model can always access one pixel, it might be sufficient to infer the values of neighboring ones, and the weights will be used to model more complicated correlations. With dropout on the inputs, the local correlations are less visible. With half the pixels missing, the model cannot rely on regularities in the input signal and should model them to make the most of each pixel. As a result, we decided to apply dropout before the lower LSTMs and after the topmost LSTM, which consistently improved the recognition results (rightmost bars in the plots of Figure 13).

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 143 of 404

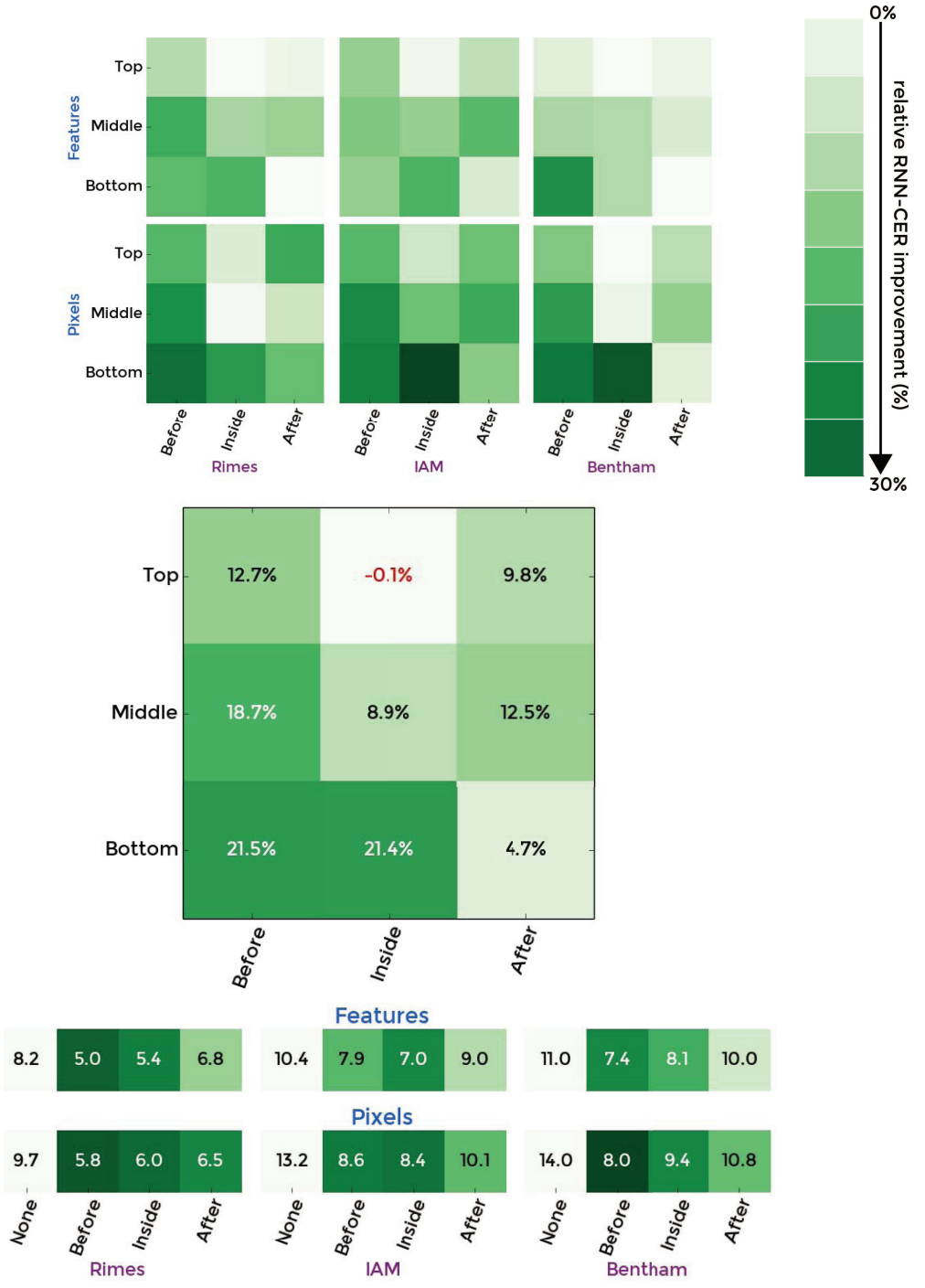


Figure 12. Effect of dropout at different positions on RNN-CER%. The relative improvement is represented by the color intensity. The top-left plot shows the result at different positions in each configuration. The top-right plot is the average. The bottom plot contains the results when dropout is applied to all layers.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 144 of 404

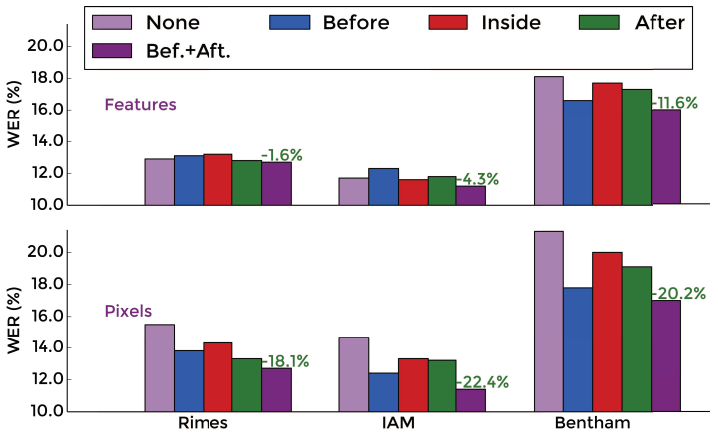


Figure 13. Effect of dropout at different positions in the complete pipeline (WER%).

6.2. Sequence-Discriminative Training

In this section, we explore the sequence-discriminative training of deep MLPs. In speech recognition, this procedure improves the results, generally by a relative 5 to 10% (Veselý et al., 2013; Su et al., 2013). Among different possibilities, we chose the state-level Minimum Bayes Risk (sMBR) criterion, described in Kingsbury (2009), which yields slightly better WERs than other sequence criteria on a speech recognition task (Switchboard; (Veselý et al., 2013)).

First, we re-aligned the training set using the cross-entropy-trained networks. Lattices are then extracted with a closed vocabulary and a language model, using the selected networks. We did not regenerate lattices during sequence training. We tried several language models, estimated from the annotations of the training set: zerogram, unigram and bigram. The zerogram is a uniform distribution over all words. We ran a few epochs of sMBR training with a learning rate of 10^{-4}.

The evolution of the WER during sMBR training is shown on Figure 14 for all databases and types of inputs. The points at epoch 0 correspond to the performance of the MLPs trained with cross-entropy. Regarding the order of the language model used to generate lattices, a zerogram, where all words have the same probability, is not sufficient: in most cases, it lead to degraded performance of the sequence-trained networks. On the other hand, a bigram language model did not yield much improvement over a unigram, and the results were even worse most of the time. With a unigram language model, for all configurations (solid lines in Figure 14), the WER was improved by sequence training.

In Figure 15, we report the results of the final systems, before and after sequence-discriminative training. We record relative WER improvements ranging from 5 to 13%, which is consistent with the observations made in speech recognition. With handcrafted features, these improvements are bigger than those observed by increasing the number of hidden layers. The success of this training procedure seems to rely on the information brought by the language model, as shown by the lack of improvement with a zerogram. However, the systems seem to benefit from the variety of the candidate sequences of words. If we increase the language constraints, changing from a unigram a bigram, the observed

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Page 145 of 404

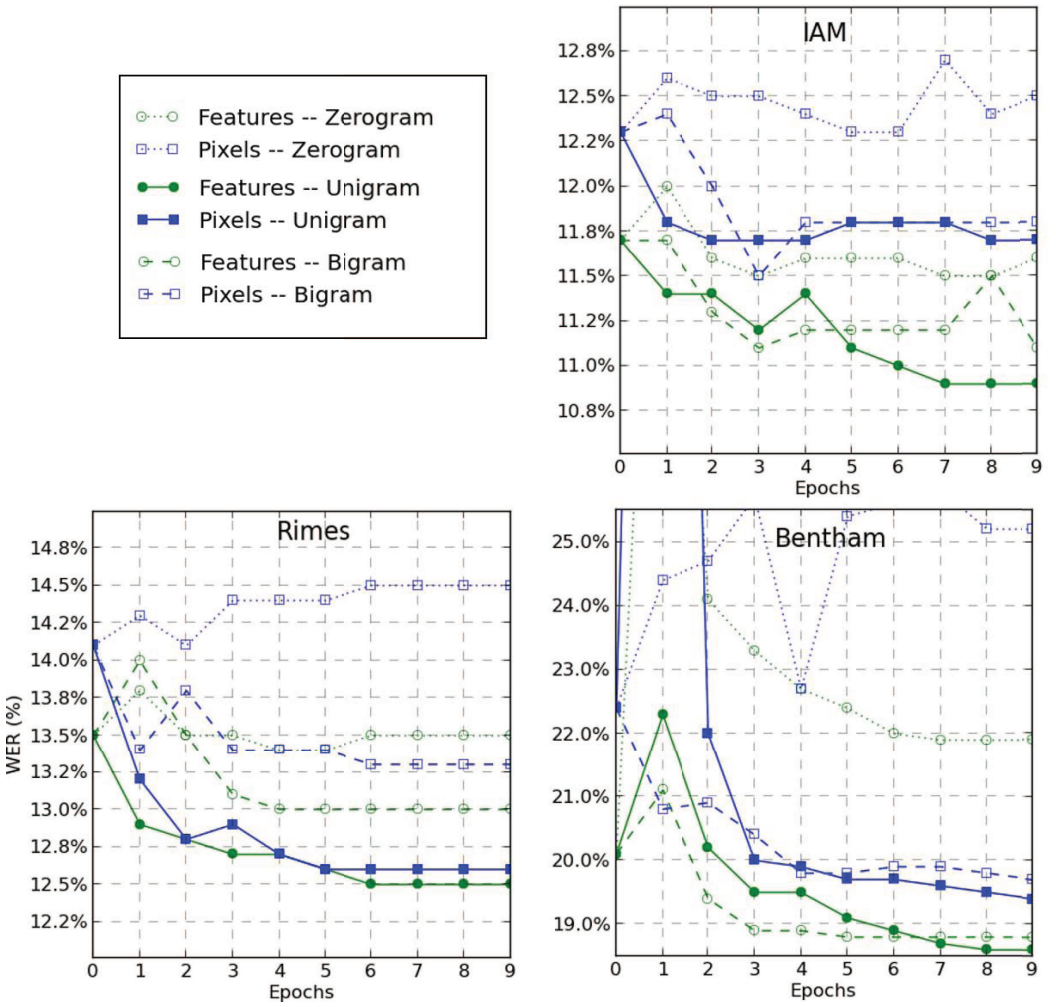


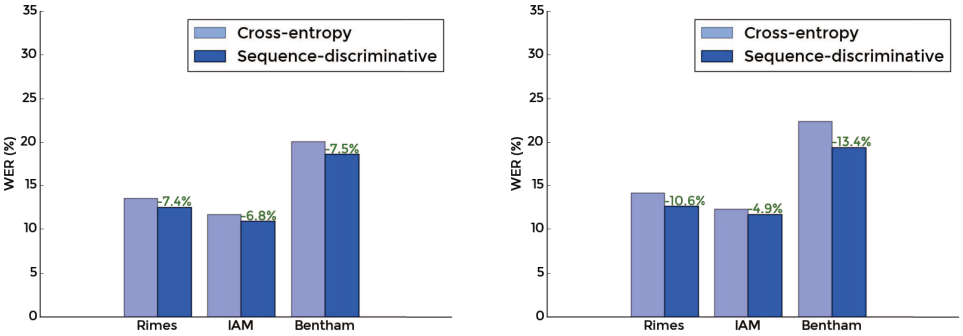
Figure 14. WER evolution during sequence-discriminative training.

improvements with respect to a cross-entropy training tend to diminish.

6.3. Framewise and CTC Training

In the previous sections, we have trained MLPs with a framewise criterion from Viterbi alignments with the baseline GMM-HMMs and RNNs with the CTC criterion to predict characters and a non-informative blank symbol. If we compare the training criteria for framewise training to predict HMM states, forward-backward training of the NN/HMM

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 146 of 404



(a) Handcrafted (b) Pixels

Figure 15. Effect of sMBR training. The *cross-entropy* corresponds to framewise training, as oposed to sMBR, which is a sequence-discriminative criterion.

system and CTC training, we minimize the following loss functions:

$$E_{Framewise}(\mathcal{S}) = -\sum_{(\mathbf{x},\mathbf{s})\in\mathcal{S}} \log \prod_t p(s_t|x_t) \tag{16}$$

$$E_{CTC}(\mathcal{S}) = -\sum_{(\mathbf{x},\mathbf{z})\in\mathcal{S}} \log \sum_{\mathbf{s}\mapsto\mathbf{z}} \prod_t p(s_t|\mathbf{x}) \tag{17}$$

$$E_{FwdBwd}(\mathcal{S}) = -\sum_{(\mathbf{x},\mathbf{z})\in\mathcal{S}} \log \sum_{\mathbf{s}\mapsto\mathbf{z}} \prod_t \frac{p(s_t|x_t)}{p(s_t)} p(s_t|s_{t-1}) \tag{18}$$

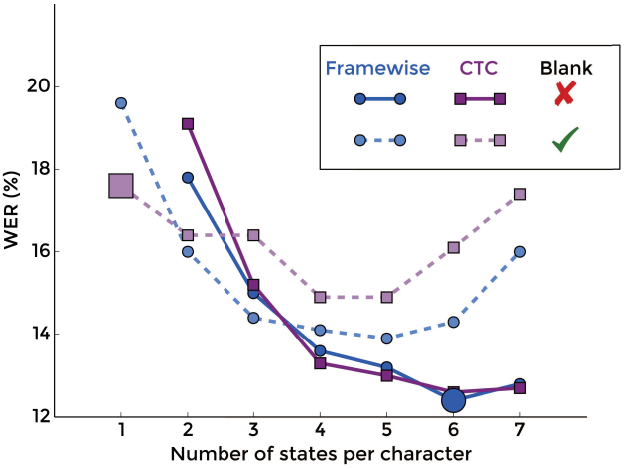
We notice that the main difference between framewise and CTC training is the summation over alternative labelings, that we also find in the forward-backward criterion. On the other hand, the main difference between CTC and forward-backward is the absence of transition and prior probabilities in the former. Hence, CTC is quite similar to HMM training without transition or prior probabilities, and with only one HMM state per character, and a "blank" state shared by all character models. It raises the question of whether it is interesting to (i) have this "blank" model, (ii) consider alternative labelings, and (iii) have only one state (or output of the network) per character.

In this section, we compare the results of framewise and CTC training of neural networks. Note that in the literature, the comparison of framewise and CTC training is carried out with the standard HMM topology with several states and no blank for framewise training, and with the CTC topology for CTC training (Graves et al., 2006; Morillot et al., 2013). Maas et al. (2014) compare CTC-trained deep neural networks with and without recurrence, using the topology defined by the CTC framework, and report considerably better results with recurrence, which we confirm in these experiments. Here, we take one step further, comparing framewise and CTC training using the same topology in each case, and observing the effect of both the training procedure and the output topology, for MLPs and RNNs.

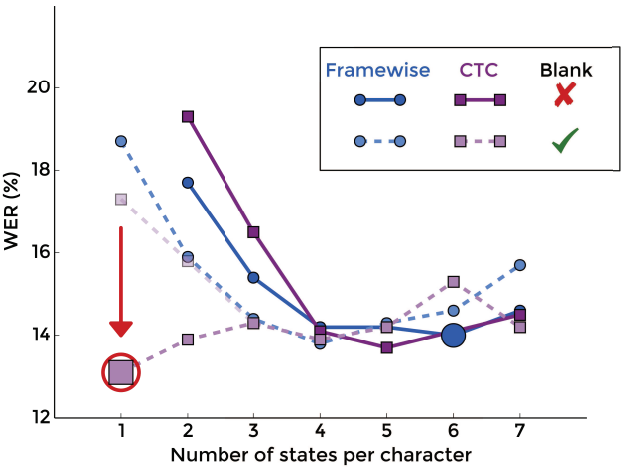
For each topology (1 to 7 states, with and without blank), we trained MLPs and RNNs. In Figure 16, we plot the WERs of MLPs (left) and RNNs (right), without blank in solid lines, and with blank in dashed ones, and using framewise training (circles) and CTC (or summation over alternatives; squares). We observe that systems with blanks are better with

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 147 of 404

a few states and worse with many states. The summation over alternative labelings does not seem to have a significant impact. Moreover, all curves but one have a similar shape: the error decreases when the number of states increases, and starts increasing when there are too many states. This increase appears sooner when we add a blank model.



(a) MLPs



(b) RNNs

Figure 16. Comparison of WER% with CTC and framewise training, with and without blank (top: MLP; bottom: RNN).

The only different case concerns the RNN with CTC training and blank symbol. The best CER is achieved with one state per character. The CTC framework, including the single state per character, blank symbol and forward-backward training is especially suited to RNNs. Moreover, CTC training without blank and with less than 5 states per character converged to a poor local optimum, for both neural networks, and most of the predictions

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 148 of 404

were whitespaces. The training algorithm did not manage to find a reasonable alignment, and the resulting WERs / CERs where above 90%. To obtain the presented results, we had to initialize the networks with one epoch of framewise training. This problem did not occur when a blank model was added, suggesting that this symbol plays a role in the success of the alignment procedure in early stages of CTC training.

7. Final Results

In the previous sections, we have carried out an evaluation of many aspects of the considered neural networks. This involved training many different neural networks and comparing the results. In particular, we have measured the impact of several modeling and design choices, leading to significant improvements. Among these networks, we selected one MLP and one RNN for each kind of inputs, achieving the best results. They are summarized in Table 8. In this section, we evaluate their performance when we optimize the complete recognition pipeline, and compare them with published results.

Table 8. Final networks selected for evaluation of the systems.

Rimes	**Features**	**Pixels**
MLP	3 hidden layers of 512 units, ± 3 context frames, sMBR training	5 hidden layers of 512 units, sMBR training
RNN	7 hidden layers of 200 units with dropout before every LSTM, CTC training	5 hidden layers of 200 units with dropout before and after every LSTM, CTC training
IAM	**Features**	**Pixels**
MLP	5 hidden layers of 256 units, ± 3 context frames, sMBR training	5 hidden layers of 1024 units, sMBR training
RNN	5 hidden layers of 200 units with dropout before the first two LSTMs and after the last one, CTC training	7 hidden layers of 200 units with dropout before and after every LSTM, CTC training

We performed the decoding with different levels of linguistic constraints. The simplest one is to recognize sequences of characters. In the next level, the lexicon is added, so that output sequences of characters form sequences of valid words. Finally, the language model is added, to promote likely sequences of words. The results are reported in Table 9. For MLPs, when the only constraint is to recognize characters, *i.e.* valid sequences of HMM states, the results are not good. The WERs are high, partly because when training the models, the recognition of a whitespace between words was optional. Therefore, the missing whitespaces in the predictions induce a high number of word merges in the output, *i.e.* a large number of deletions and substitutions. When a vocabulary is added, the error rates are roughly divided by two. Another reduction by a factor two is achieved when a language model is present. These results show the importance of the linguistic constraints to correct the numerous errors of the MLP/HMM system.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 149 of 404

Table 9. Effect of adding linguistic knowledge in NN/HMM systems.

	MLP-Features		MLP-Pixels		RNN-Features		RNN-Pixels	
	WER%	CER%	WER%	CER%	WER%	CER%	WER%	CER%
Rimes								
no lexicon	61.1	17.8	59.5	17.8	20.1	5.1	20.9	5.6
lexicon	26.9	6.8	26.1	7.2	16.7	5.3	16.4	4.3
lexicon+LM	12.5	3.4	12.6	3.8	12.8	3.8	12.7	4.0
IAM								
no lexicon	54.7	15.8	54.2	15.6	27.5	7.9	24.7	7.3
lexicon	24.7	7.7	25.5	8.0	17.6	5.5	16.7	5.3
lexicon+LM	10.9	3.7	11.7	4.0	11.2	3.8	11.4	3.9

For RNNs, we notice that the differences between no constraints and lexicon with LM are not as dramatic as for MLPs. The WERs are only multiplied by 2 to 2.5 when we remove the constraints, when it was roughly multiplied by 5 for MLPs. As mentioned previously, a lot of context is used by the network through the recurrent connections, which seems to enable the network to predict characters with some knowledge about the words. Yet, both the lexicon and the language model bring significant improvements, and remain very important to achieve state-of-the-art results. The fact that the RNNs produce reasonably good transcriptions by themselves should make them more suited to open-vocabulary scenarios (*e.g.* the approaches of (Kozielski et al., 2013b; Messina and Kermorvant, 2014)), where the language model is either at the character level, or a hybrid between a word and a character language model.

The final results, comparing different models and input features, and comparing our proposed systems with other published results, are reported in Tables 10 (Rimes) and 11 (IAM). The error rates are reported on both the validation and evaluation sets. The conclusions of the previous sections about the small differences in performance between MLPs and RNNs and between features and pixels are still applicable to the evaluation set results.

The systems based on neural networks outperform the GMM-HMM baseline systems: the relative improvement is about 30%. Moreover, on Rimes, we see that all of our single systems achieve state of the art performance, competing with the systems of Pham et al. (2014), which uses the same language model with an MDLSTM-RNN with dropout, trained directly on the image, and of Doetsch et al. (2014), a hybrid BLSTM-RNN. On IAM, it is worth noting that the decoders of Kozielski et al. (2013a) and Doetsch et al. (2014) include an open-vocabulary language model which can potentially recognize any word, when the error of our systems is bound to be higher than the OOV rate of 3.7%. For Kozielski et al. (2013a), the second result in Table 11 corresponds to the closed vocabulary decoding with the same system as the first one. Unfortunately, the results on the evaluation set are not reported with this setup, but from the validation set errors, we may consider that our single systems achieve similar performance as the best closed-vocabulary systems of Pham et al. (2014) and Kozielski et al. (2013a).

For each database, we have selected four systems, two MLPs and two RNNs, with feature and pixel inputs. We have seen that their performance was comparable. However, the

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 150 of 404

Table 10. Final results on Rimes database

		Dev.		Eval.	
		WER%	**CER%**	**WER%**	**CER%**
GMM-HMM	Features	17.2	5.9	15.8	6.0
MLP	Features	12.5	3.4	12.7	3.7
	Pixel	12.6	3.8	12.4	3.9
RNN	Features	12.8	3.8	12.6	3.9
	Pixels	12.7	4.0	13.8	4.6
ROVER combination		11.3	3.5	11.3	3.7
Lattice combination		11.2	3.3	**11.2**	3.5
(Pham et al., 2014)		-	-	12.3	**3.3**
(Doetsch et al., 2014)		-	-	12.9	4.3
(Messina and Kermorvant, 2014)		-	-	13.3	-
(Kozielski et al., 2013a)		-	-	13.7	4.6
(Messina and Kermorvant, 2014)		-	-	14.6	-
(Menasri et al., 2012)		-	-	15.2	7.2

Table 11. Final results on IAM database

		Dev.		Eval.	
		WER%	**CER%**	**WER%**	**CER%**
GMM-HMM	Features	15.2	6.3	19.6	9.0
MLP	Features	10.9	3.7	13.3	5.4
	Pixel	11.4	3.9	13.8	5.6
RNN	Features	11.2	3.8	13.2	5.0
	Pixels	11.8	4.0	14.4	5.7
ROVER combination		9.6	3.6	11.2	4.7
Lattice combination		9.6	3.3	**10.9**	**4.4**
(Doetsch et al., 2014)		**8.4**	**2.5**	12.2	4.7
(Kozielski et al., 2013a)		9.5	2.7	13.3	5.1
(Pham et al., 2014)		11.2	3.7	13.6	5.1
(Kozielski et al., 2013a)		11.9	3.2	-	-
(Messina and Kermorvant, 2014)		-	-	19.1	-
(Espana-Boquera et al., 2011)		19.0	-	22.4	9.8

differences between these systems probably lead to different errors. Thus, we combined their outputs, with two methods: ROVER (Fiscus, 1997), which combines the transcription outputs, and a lattice combination technique (Xu et al., 2011), which extracts the final transcript from the combination of lattice outputs. For both methods, we started by computing the decoding lattices, obtained with the decoder implemented in Kaldi. As one can see in Tables 10 and 11, both combination methods clearly outperform the best published WERs on Rimes and IAM, even those obtained with open-vocabulary systems.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 151 of 404

Conclusion

In this chapter, we focused on the problem of offline handwritten text recognition, consisting of transforming images of cursive text into their digital transcription. More specifically, we concentrated on images of text lines, and we adopted the popular sliding window approach: a sequence of feature vectors is extracted from the image, processed by an optical model, and the resulting sequence is modeled by Hidden Markov Models and linguistic knowledge (a vocabulary and a language model) to obtain the final transcription. In the interest of gaining a deeper knowledge or understanding of these models, we have carried out thorough experiments with deep neural network optical models for hybrid NN/HMM handwriting recognition. We focused on two popular architectures: Multi-Layer Perceptrons, and Long Short-Term Memory Recurrent Neural Networks. We investigated many aspects of those models: the type of inputs, the output model, the training procedure, and the architecture of the networks. We answered the following questions regarding neural network optical models.

⟶ *Is it still important to design handcrafted features when using deep neural networks, or are pixel values sufficient?*

Although we have seen that shallow networks tend to be much better when fed with handcrafted features, we showed that the discrepancy between the performance of the systems with handcrafted feature and pixel inputs is largely decreased with deep neural networks. This supports the idea that an automatic extraction of learnt features happen in the lower layers of the network. Neural networks with pixel inputs require more hidden layers, but finally achieve similar performance as networks operating with handcrafted features. The need to design and implement good feature extractions may therefore not be necessary.

⟶ *Can deep neural networks give rise to big improvements over neural networks with one hidden layer for handwriting recognition?*

We have trained two kinds of neural networks, namely Multi-Layer Perceptrons and Recurrent Neural Networks, and we have evaluated the influence of the number of hidden layers on the performance of the system. We trained neural networks of different depths, and we have shown that deep neural networks achieve significantly better results than neural networks with a single hidden layer. With deep neural networks, we recorded relative improvements of error rates in the range 5-10% for MLPs and 10-15% for RNNs. When the inputs of the network are pixels, the improvement can be much larger.

⟶ *What are the important characteristics of Recurrent Neural Networks, which make them so appropriate for handwriting recognition?*

We have seen that explicitly including context in the observation sequences did not improve the results, as it does for MLPs, and that RNNs could effectively learn the dependencies in the input sequences, and the context necessary to make character predictions. We have shown that the recurrence was especially useful in the top layers of RNNs, at least in the CTC framework. We have also shown that RNNs can take advantage of the CTC framework, which defines an objective function at the sequence level for training, but also the output classes of the network. These are characters directly, and a special *non-character* symbol, allowing the network to produce transcriptions with the neural network alone, without relying on an HMM or any other elaborated model.

⟶ *How (deep) Multi-Layer Perceptrons compare to the very popular Recurrent Neu-*

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 152 of 404

ral Networks, which are now widespread in handwriting recognition and achieve state-of-the-art performance?

We have shown that deep MLPs can achieve similar performance to RNNs, and that both kinds of model give comparable results to the state-of-the-art on Rimes and IAM. We conclude that despite the dominance of RNNs in the literature of handwriting recognition, MLPs, and possibly other kinds of models, can be a good alternative, and therefore should not be put aside. However, we have also shown that MLPs are more sensitive to the number of states in HMM models, and to the amount of input context provided. The RNNs, with CTC training, model sequences of characters directly, and are much easier to train, coping with the input sequence and the length estimation automatically.

⟶ *What are the good training strategies for Neural Networks for handwriting recognition? Can the Connectionist Temporal Classification paradigm be applied to other Neural Networks? What improvements can be observed with a discriminative criterion at the sequence level?*

The optimized cost is an important feature of the training procedure of models with machine learning algorithms and it may affect the quality of the system. The most common approach to training neural networks for hybrid NN/HMM systems consists in first aligning the frames to HMM states with a bootstrapping system, and then train the network on the obtained a labeled dataset with a framewise classification cost function, such as the cross-entropy. This strategy amounts to considering the segmentation of the input sequence into HMM states fixed, and to have the network predict it. A softer approach, similar to the Baum-Welch training algorithm, would consist in summing over all possible segmentations of the input sequences yielding the same final transcription. We have seen that in general, this approach produces only small improvements.

The CTC framework is such a training procedure but also defines the outputs of the neural network to correspond to the set of characters, and a special *non-character* output (*blank* label). We have shown that RNNs can achieve good results with the CTC criterion. MLPs can be trained with CTC but do not benefit from it.

We have studied the effects of applying a discriminative training criterion at the sequence level, namely state-level Minimum Bayes Risk (sMBR). We have shown that fine-tuning the MLPs with sMBR yields significant improvements, between 5 and 13% of WER, which is consistent with the speech recognition literature. Moreover, we investigated a new regularization technique, dropout, in RNNs, extending the work of (Pham et al., 2014; Zaremba et al., 2014). We reported significant improvements over the method presented in Pham et al. (2014) when dropout is applied before LSTM layers rather than after them.

Finally, all our models achieved error rates comparable to the state-of-the-art on Rimes and IAM, independently of the type of inputs (handcrafted features or pixels), and of the kind of neural network (MLP or RNN). The lattice combination of our systems, with the method of Xu et al. (2011), outperformed the best published systems for all three databases, showing the complementarity of the developed models.

References

Augustin, E., Carré, M., Grosicki, E., Brodin, J.-M., Geoffrois, E., and Preteux, F. (2006). RIMES evaluation campaign for handwritten mail processing. In *Proceedings of the*

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 153 of 404

Workshop on Frontiers in Handwriting Recognition, number 1.

Bengio, Y., De Mori, R., Flammia, G., and Kompe, R. (1992). Global optimization of a neural network-hidden Markov model hybrid. *Neural Networks, IEEE Transactions on*, 3(2):252–259.

Bertolami, R. and Bunke, H. (2008). Hidden Markov Model Based Ensemble Methods for Offline Handwritten Text Line Recognition. *Pattern Recognition*, 41(11):3452 – 3460.

Bianne, A.-L., Menasri, F., Al-Hajj, R., Mokbel, C., Kermorvant, C., and Likforman-Sulem, L. (2011). Dynamic and Contextual Information in HMM modeling for Handwriting Recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(10):2066 – 2080.

Bianne-Bernard, A.-L. (2011). *Reconnaissance de mots manuscrits cursifs par modèles de Markov cachés en contexte*. PhD thesis, Telecom ParisTech.

Bloomberg, D. S., Kopec, G. E., and Lakshmi Dasari (1995). Measuring document image skew and orientation. *Proc. SPIE Document Recognition II*, 2422(302):302–316.

Bluche, T., Louradour, J., Knibbe, M., Moysset, B., Benzeghiba, M. F., and Kermorvant, C. (2014). The A2iA Arabic Handwritten Text Recognition System at the Open HaRT2013 Evaluation. In *11th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 161–165. IEEE.

Bourlard, H. and Morgan, N. (1994). *Connectionist speech recognition: a hybrid approach Chapter 7*, volume 247 of *The Kluwer international series in engineering and computer science: VLSI, computer architecture, and digital signal processing*. Kluwer Academic Publishers.

Bourlard, H. and Wellekens, C. J. (1989). Links between Markov models and multilayer perceptrons. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(12):1167–1178.

Bridle, J. S. (1990a). Alpha-nets: a recurrent "neural" network architecture with a hidden Markov model interpretation. *Speech Communication*, 9(1):83–92.

Bridle, J. S. (1990b). Probabilistic interpretation of feedforward classification network outputs with relationships to statistical pattern recognition. In *Neurocomputing*, pages 227–236. Springer.

Buse, R., Liu, Z. Q., and Caelli, T. (1997). A structural and relational approach to handwritten word recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 27(5):847–61.

Doetsch, P., Kozielski, M., and Ney, H. (2014). Fast and robust training of recurrent neural networks for offline handwriting recognition. pages –.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

El-Hajj, R., Likforman-Sulem, L., and Mokbel, C. (2005). Arabic handwriting recognition using baseline dependant features and hidden markov modeling. In *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*, pages 893–897. IEEE.

Espana-Boquera, S., Castro-Bleda, M. J., Gorbe-Moya, J., and Zamora-Martinez, F. (2011). Improving offline handwritten text recognition with hybrid HMM/ANN models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(4):767–779.

Fiscus, J. G. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU1997)*, pages 347–354. IEEE.

Gers, F. (2001). Long Short-Term Memory in Recurrent Neural Networks TH ESE. 2366.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *International Conference on Machine learning*, pages 369–376.

Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., and Schmidhuber, J. (2009). A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–68.

Graves, A., Mohamed, A.-R., and Hinton, G. (2013). Speech Recognition with Deep Recurrent Neural Networks. In *proc ICASSP*, number 3.

Graves, A. and Schmidhuber, J. (2008). Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks. In *Advances in Neural Information Processing Systems*, pages 545–552.

Grosicki, E. and El-Abed, H. (2011). Icdar 2011-french handwriting recognition competition. In *International Conference on Document Analysis and Recognition (ICDAR2011)*, pages 1459–1463. IEEE.

Haffner, P. (1993). Connectionist speech recognition with a global MMI algorithm. In *EUROSPEECH*.

Hennebert, J., Ris, C., Bourlard, H., Renals, S., and Morgan, N. (1997). Estimation of global posteriors and forward-backward training of hybrid HMM/ANN systems.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Janet Holmes, B. V. and Johnson, G. (1998). Guide to the wellington corpus of spoken new zealand english.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 155 of 404

Johansson, S. (1980). The LOB corpus of British English texts: presentation and comments. *ALLC journal*, 1(1):25–36.

Kingsbury, B. (2009). Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)*, pages 3761–3764. IEEE.

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181–184. IEEE.

Konig, Y., Bourlard, H., and Morgan, N. (1996). Remap: Recursive estimation and maximization of a posteriori probabilities-application to transition-based connectionist speech recognition. *Advances in Neural Information Processing Systems*, pages 388–394.

Kozielski, M., Doetsch, P., Hamdani, M., and Ney, H. (2014). Multilingual Off-line Handwriting Recognition in Real-world Images. pages 1–1.

Kozielski, M., Doetsch, P., Ney, H., et al. (2013a). Improvements in RWTH's System for Off-Line Handwriting Recognition. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 935–939. IEEE.

Kozielski, M., Rybach, D., Hahn, S., Schluter, R., and Ney, H. (2013b). Open vocabulary handwriting recognition using combined word-level and character-level language models. In *38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2013)*, pages 8257–8261. IEEE.

Maas, A. L., Hannun, A. Y., Jurafsky, D., and Ng, A. Y. (2014). First-Pass Large Vocabulary Continuous Speech Recognition using Bi-Directional Recurrent DNNs. *arXiv preprint arXiv:1408.2873*.

Marti, U.-V. and Bunke, H. (2002). The IAM-database: an English sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46.

Menasri, F., Louradour, J., Bianne-Bernard, A.-L., and Kermorvant, C. (2012). The A2iA French handwriting recognition system at the Rimes-ICDAR2011 competition. In *Document Recognition and Retrieval Conference*, volume 8297.

Messina, R. and Kermorvant, C. (2014). Surgenerative Finite State Transducer n-gram for Out-Of-Vocabulary Word Recognition. In *11th IAPR Workshop on Document Analysis Systems (DAS2014)*, pages 212–216.

Mohri, M., Pereira, F., and Riley, M. (2002). Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16(1):69–88.

Morillot, O., Likforman-Sulem, L., and Grosicki, E. (2013). Comparative study of HMM and BLSTM segmentation-free approaches for the recognition of handwritten text-lines. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 783–787. IEEE.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 156 of 404

Moysset, B., Bluche, T., Knibbe, M., Benzeghiba, M. F., Messina, R., Louradour, J., and Kermorvant, C. (2014). The A2iA Multi-lingual Text Recognition System at the second Maurdor Evaluation. In *14th International Conference on Frontiers in Handwriting Recognition (ICFHR2014)*, pages 297–302.

Otsu, N. (1979). A Threshold Selection Method from Grey-Level Histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66.

Pesch, H., Hamdani, M., Forster, J., and Ney, H. (2012). Analysis of Preprocessing Techniques for Latin Handwriting Recognition. *ICFHR*, 12:18–20.

Pham, V., Bluche, T., Kermorvant, C., and Louradour, J. (2014). Dropout improves recurrent neural networks for handwriting recognition. In *14th International Conference on Frontiers in Handwriting Recognition (ICFHR2014)*, pages 285–290.

Povey, D. (2004). *Discriminative training for large vocabulary speech recognition*. Ph.D. thesis, Cambridge University.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The Kaldi speech recognition toolkit. In *Workshop on Automatic Speech Recognition and Understanding (ASRU2011)*, pages 1–4.

Renals, S., Morgan, N., Bourlard, H., Cohen, M., and Franco, H. (1994). Connectionist probability estimators in HMM speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 2(1):161–174.

Roeder, P. (2009). *Adapting the rwth-ocr handwriting recognition system to french handwriting*. Master's thesis, Human Language Technology and Pattern Recognition Group, RWTH Aachen University, Aachen. Germany.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1988). Learning representations by back-propagating errors. *Cognitive modeling*.

Sainath, T. N., Mohamed, A.-r., Kingsbury, B., and Ramabhadran, B. (2013). Deep convolutional neural networks for LVCSR. In *38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2013)*, pages 8614–8618. IEEE.

Sánchez, J. A., Mühlberger, G., Gatos, B., Schofield, P., Depuydt, K., Davis, R. M., Vidal, E., and de Does, J. (2013). tranScriptorium: a european project on handwritten text recognition. In *Proceedings of the 2013 ACM symposium on Document engineering*, pages 227–228. ACM.

Sánchez, J. A., Romero, V., Toselli, A., and Vidal, E. (2014). ICFHR 2014 HTRtS: Handwritten Text Recognition on tranScriptorium Datasets. In *International Conference on Frontiers in Handwriting Recognition (ICFHR)*.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 157 of 404

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45(11):2673–2681.

Senior, A. and Robinson, T. (1996). Forward-backward retraining of recurrent neural networks. *Advances in Neural Information Processing Systems*, pages 743–749.

Stolcke, A. (2002). SRILM – An Extensible Language Modeling Toolkit. In *International Conference on Spoken Language Processing*.

Su, H., Li, G., Yu, D., and Seide, F. (2013). Error back propagation for sequence training of Context-Dependent Deep Networks for conversational speech transcription. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2013)*, pages 6664–6668.

Toselli, A. H., Juan, A., González, J., Salvador, I., Vidal, E., Casacuberta, F., Keysers, D., and Ney, H. (2004). Integrated handwriting recognition and interpretation using finite-state models. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(04):519–539.

Toselli, A. H., Romero, V., Pastor, M., and Vidal, E. (2010). Multimodal interactive transcription of text images. *Pattern Recognition*, 43(5):1814–1825.

Veselý, K., Ghoshal, A., Burget, L., and Povey, D. (2013). Sequence-discriminative training of deep neural networks. In *14th Annual Conference of the International Speech Communication Association (INTERSPEECH2013)*, pages 2345–2349.

Vinciarelli, A. and Luettin, J. (2001). A new normalisation technique for cursive handwritten words. *Pattern Recognition Letters*, 22:1043–1050.

W. N. Francis, H. K. (1979). Brown corpus manual.

Xu, H., Povey, D., Mangu, L., and Zhu, J. (2011). Minimum Bayes Risk decoding and system combination based on a recursion for edit distance. *Computer Speech & Language*, 25(4):802–828.

Yan, Y., Fanty, M., and Cole, R. (1997). Speech recognition using neural networks with forward-backward probability generated targets. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 4, pages 3241–3241. IEEE Computer Society.

Zaremba, W., Sutskever, I., and Vinyals, O. (2014). Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 158 of 404

In: Handwriting: Recognition, Development and Analysis ISBN: 978-1-53611-937-4
Editors: Byron L. D. Bezerra et al. © 2017 Nova Science Publishers, Inc.

Chapter 6

HANDWRITTEN AND PRINTED IMAGE DATASETS: A REVIEW AND PROPOSALS FOR AUTOMATIC BUILDING

***Gearlles V. Ferreira*[1,*], *Felipe M. Gouveia*[1,†], *Byron L. D. Bezerra*[1,‡], *Eduardo Muller*[1,§], *Cleber Zanchettin*[2,¶] *and Alejandro Toselli*[3,‖]**
[1]E-Comp, Universidade de Pernambuco, Recife, Brazil
[2]Centro de Informática
Universidade Federal de Pernambuco, Recife, Brazil
[3]Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València, València, Spain

1. Introduction

The use of the database is of fundamental importance for pattern recognition processes, supervised training and computing in general (Duda et al., 2012). For problems related to cursive handwriting and optical character recognition it is not different. Over the years a great effort has been developed by the scientific community with the goal to develop these datasets (Yalniz and Manmatha, 2011; Lazzara and Géraud, 2014; Padmanabhan et al., 2009; Fischer et al., 2012, 2011a; Marti and Bunke, 1999; Shahab et al., 2010). The construction process of these however is not very automated, and involves a great effort by the researcher (Marti and Bunke, 1999). Usually, there are two approaches to build datasets, which are distinguished by documents are captured: natural datasets and artificial datasets.

Natural datasets are the one built from the scanning and processing of actual real documents. These datasets normally have a more real challenge but are more difficult to process

*E-mail address: gvf@ecomp.poli.br.
†E-mail address: fmg@ecomp.poli.br.
‡E-mail address: byronleite@ecomp.poli.br.
§E-mail address: eam@ecomp.poli.br.
¶E-mail address: cz@cin.ufpe.br.
‖E-mail address: ahector@prhlt.upv.es.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

due to the variety of the nature of the documents. Below we discuss some existing natural datasets.

On the other hand, artificial datasets are built from the use of forms or applications for assistance to third, making that the texts and elements of the datasets are not derived from actual real documents. In this category we have the datasets like IamDB (Marti and Bunke, 1999), Rimes (Menasri et al., 2012) and CVL (Kleber et al., 2013).

In this chapter, we review and analyze the most used handwritten and printed text datasets developed during the last few decades. We not only discuss the dataset format, but also its structure, category, statistics and how they are used in the literature. The datasets are extremely important for machine learning algorithms and in a review of the most advances in the field some researchers suggest a provocative explanation: perhaps many major machine learning breakthroughs have actually been constrained by the availability of high-quality training datasets, and not by algorithmic advances[1]. Considering this importance to machine learning and consequently to handwriting recognition advances, we discuss the complexity of building datasets and present two techniques to generate datasets for both handwritten and printed texts.

2. Dataset Review

There is a large number of handwritten and printed text datasets that have been used for different purposes over the last years. These datasets could be categorized by different aspects, for example, data acquisition type (online or offline), text type (handwritten or printed), size, format, tasks supported, language and others. This section presents a detailed discussion and review about the handwritten and printed text datasets developed in each of those aspects.

2.1. NIST Handprinted Forms and Characters Database

Compiled by the National Institute of Standards and Technology the database contains handprinted sample forms from 3600 writers, 810,000 character images isolated from their forms, ground truth classifications for those images, reference forms for further data collection, and software utilities for image management and handling (Garris et al., 1997). The main database is divided into different datasets: **NIST Handprinted Forms and Characters** NIST Special Database 19 contains NIST's entire corpus of training materials for handprinted document and character recognition; **NIST Machine-Print Database of Gray Scale and Binary Images (MPDB)** The NIST machine-printed database (Special Database 8) contains gray scale and binary images of machine printed pages. There is a total of 3,063,168 characters in the set. A reference file is included for each page; **NIST Structured Forms Reference Set of Binary Images II (SFRS2)** The second NIST database of structured forms (Special Database 6) consists of 5,595 pages of binary, black-and-white images of synthesized documents containing hand-print. The documents in this database are 12 different tax forms with the IRS 1040 Package X for the year 1988.

[1] Alexander Wissner-Gross

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 160 of 404

THE MNIST DATABASE of handwritten digits proposed by LeCun et al. (2012), has a training set of 60,000 examples[2], and a test set of 10,000 examples. It is a subset of a larger set available from NIST. The digits have been size-normalized and centered in a fixed-size image.

2.2. IAM Datasets

The Institute of Informatics and Applied Mathematics (IAM) at the University of Bern, Switzerland has been involved in the development of a large amount of handwritten datasets. There are four main datasets: IAM Handwriting Database, IAM On-Line Handwriting Database, IAM Online Document Database and IAM Historical Document Database, but the last one is divided into three smaller datasets (Saint Gall Database, Parzival Database and Washington Database). Each dataset will be discussed in the following paragraphs.

IAM Handwriting Database. The IAM Handwriting Database (Marti and Bunke, 1999) was published in 1999 and is formed by English text forms. The dataset is mainly used for handwritten recognition, writer identification and verification tasks. 657 writers contributed with samples of their writing, resulting in 1,539 pages of scanned text, 5,685 sentences, 13,353 text lines and 115,320 words. Each form was scanned at a resolution of 300 dpi and the images were stored in eps format using 256 gray levels, each one linked to a XML file containing meta information (e.g. the labels).

IAM On-Line Handwriting Database. The IAM On-Line Handwriting Database (IAM-OnDB) (Liwicki and Bunke, 2005b) was published at the ICDAR 2005 and contains forms of handwritten English sentences acquired on a whiteboard. The dataset is stored in a XML containing writer-id, writer gender, writer native language and the transcription. It contains 86,272 words with 13,049 text lines written by 221 writers. The transcription are available in TIFF format and was used for recognition purposes (Liwicki and Bunke, 2005a, 2006) and writer identification (Schlapbach et al., 2008).

IAM Online Document Database. The IAM Online Document Database (Indermühle et al., 2010) contains 941 online handwritten and printed documents (diagrams, drawings, tables, lists, text blocks, formulas and markings) acquired with a digital pen. The dataset provides the collected metadata in XML format and has been used for different tasks such as handwritten text recognition, document layout analysis. The IAM Online Document Database consists of approximately 70,000 words and more than 7,500 text lines.

Saint Gall Database. The Saint Gall Database (Fischer et al., 2011a) is a historical dataset from a single writer using ink on parchment. In the Latin language from the 9th century, the dataset includes 60 pages corresponding 1,410 text lines, 11,597 words, 4,890 word labels and 49 unique letters. The original manuscript which the database was scanned is housed at the Abbey Library of Saint Gall, Switzerland. The manuscript was scanned at

[2]http//yann.lecun.com/exdb/mnist/

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 161 of 404

300 dpi on JPEG format and was pre-processed using binarization and normalization operations. The Saint Gall Database has been employed for text lines and words segmentation as well as handwritten recognition (Fischer et al., 2010a).

Parzival Database. The Parzival Database (Fischer et al., 2009) is a historical dataset published in 2009 containing a 13th century manuscript in medieval German language composed by three writers. The original manuscript was written using ink on parchment and, like the Saint Gall Database, is housed at the Abbey Library of Saint Gall, Switzerland. The dataset has 47 pages including 4,477 lines, 23,478 words and 93 unique letters. It's divided into pages images (JPEG, 300 dpi) after binarization and normalization operations. The Parzival Database has been used for text line segmentation (Fischer et al., 2012, 2011b) and word recognition (Fischer et al., 2009, 2010b).

Washington Database. The Washington Database (Rath and Manmatha, 2007a) was created from the George Washington Papers at the Library of Congress. The dataset contains 18th century English words alongside with their transcription. It includes 20 pages of 656 text lines, 4,894 words and 82 unique letters. All images were binarized and normalized. This dataset is mainly used for word-level recognition (Fischer et al., 2012; Frinken et al., 2012) and keyword spotting (Fischer et al., 2013; Rath and Manmatha, 2007b).

2.3. Bentham

The Betham dataset (Gatos et al., 2014) is a large set of scanned documents written in the 18th century by the philosopher Jeremy Bentham. The dataset was built using a crowded-funded web platform where volunteers help transcribing the documents. There are more than 6,000 documents and it's provided in two parts: the images and the ground-truth. The latter has information about the layout and the transcription for each line of the documents. This dataset was used for a competition in ICFHR 2014 for handwritten text recognition.

2.4. RIMES

The RIMES dataset was designed focused on recognition of handwritten letters sent by customers via postal mail to companies. To build the database, 1300 people have participated writing up to 5 mails. Each email contains two to three pages resulting in 12,723 pages and a total of 5605 mails. The RIMES database has been used for several competitions (ICFHR 2008, ICDAR 2009, ICDAR 2011) and is used for different tasks such as handwritten recognition and mail classification (Kermorvant and Louradour, 2010).

2.5. Maurdor

The Maurdor database (Brunessaux et al., 2014) was published in 2013 by the French National Metrology and Testing Laboratory with both handwritten and printed text. The dataset contains a total of 2,500 documents in English, French and Arabic and in different types (forms, business documents, letters, correspondences and newspaper articles). This database was been used for a competition organized by the Laboratoire National de

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 162 of 404

Métrologie (LEN) with tasks in zone area segmentation, writing type identification, optical character recognition and logical structure extraction.

2.6. PRImA

The PRImA database (Antonacopoulos et al., 2009) is a printed text dataset containing realistic documents with a variety of layouts, types, structure and font. The database was built by scanning magazines, technology publications and technical articles resulting in 305 ground-truthed images. In addition to the images, the dataset provides searchable document-level metadata and a web interface for navigation. Since the PRImA dataset has documents with a variety of layouts, it was initially used for layout analysis tasks, but it has been used for optical character recognition (Diem et al., 2011) as well.

3. Handwriting Synthesis

Cursive handwriting synthesis is aimed at generating either a handwritten text image or a pen path information involving online handwriting trajectories. Any of these outputs (image/pen path information) is characterized by trying to look as much as possible to the handwriting of real people. One of the main motivations for developing these techniques is to help to expand already-existing datasets for handwriting recognition, overcoming the difficulty and time it takes to create one from scratch (Elarian et al., 2015; Marti and Bunke, 1999).

Handwriting synthesis is an old research area with works dating from 1996 (Guyon, 1996), and the most recent advances in handwriting recognition models, such as Graves (2012) and Toselli and Vidal (2015) along with the use of deep learning techniques (Pham et al., 2014) has increased the necessity of expanding the existing datasets and thereby leading to a renewed interest in handwriting synthesis (Ahmad and Fink, 2015; Dinges et al., 2015; Chen et al., 2015). These works and their contributions to the field will be examined bellow, classifying them into three main areas: symbols and connections, statistical and machine learning.

3.1. Symbols and Connections

Techniques based on the use of symbols and connections among them to generate new writing samples are perhaps the oldest form of synthesis as can be seen in Guyon (1996). Advances in this area have produced interesting results and contributed to the improvement of classification systems as in (Jawahar and Balasubramanian, 2006; Al-Muhtaseb et al., 2011; Elarian et al., 2015).

In the work “An Arabic handwriting synthesis system” Elarian et al. (2015) presents a new synthesis model based on characters and connections. In this work, with the help of volunteers, a dataset of handwritten text which covers all possible Arabic character-shapes has been developed to be used as a baseline system. From this dataset, previously segmented characters were selected and connections among them were properly performed to synthesize required words. Experimental results were obtained on the dataset IFN/ENIT (Pechwitz et al., 2002) using the HTK Toolkit (Young et al., 2002), which implements a handwrit-

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 163 of 404

ten text recognition based on Hidden Markov Models (HMMs). The final system evaluation was carried out for different numbers of generated word samples: 1 to 12. Obtained results are reported in Table 1.

Table 1. Results obtained by Elarian et al. (2015)

Samples	Result		
	Top 1	Top 5	Top 10
Baseline system	48.52	64.17	67.74
One sample	64.51	78.09	81.67
Six samples	70.13	82.94	85.53
Twelve samples	70.58	84.22	87.03

As can be observed in Table 1, the addition of synthesized word samples results in an increase of the classification rate, which confirms the effectiveness of the proposed method. However, this approach mainly involves two drawbacks that need to be taken into account: the creation of a dataset for the language/alphabet desired, and the required manual segmentation of it into the corresponding symbols/characters. The problem of dataset creation can ony be solved partially when the demand for generated synthetic data has a growth trend and hence requiring a larger number of samples in the input alphabet.

3.2. Statistical

Some models of synthesis to generate artificial text are based on statistical knowledge about how the actual text is produced. To collect such a statistical information, it is more common to use datasets based on online handwriting, as in Martín-Albo et al. (2014) and Plamondon et al. (2014).

Martin et al. in "Training of On-line Handwriting Text Recognizers with Synthetic Text Generated Using the Kinematic Theory of Rapid Human Movements" Martín-Albo et al. (2014) describes a synthesis model based on human kinematic motion. In this paper, the rational principle behind the proposed approach is that the response to a given gesture is modeled by two log-normal distributions: one modeling the agonist in the same direction of the gesture, and another modeling the antagonist in the opposite direction of the gesture. For complex movements, such as writing, movement can be modeled by a vector sum of log-normals. The generation of the synthesis model consists of three phases: first the signal parameters are extracted from some online data; second noise generation is carried out upon this data, and third, speed is adjusted and a new sequence of coordinates pairs (x, y) is generated. For the experiments, the single handwriting word dataset "Unipen-ICROW-03" was employed, and an on-line handwritten text recognition based on left-to-right HMMs with a variable number of states and 8 Gaussians per state. In this case, for each word and author available in the dataset, s new synthetic words were generated. The final obtained results are summarized in Table 2, which shows the classification error rate for each combination of a number of authors-specific samples and number of synthetically generated samples (s).

The results reported by Martin et al. are very promising, both for the technical simplicity and for the ease of use thereof. However, this approach depends strongly on the dataset

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 164 of 404

Table 2. Results reported by Martín-Albo et al. (2014)

# Author	Samples					
	10	20	50	100	150	200
20	11.0	10.40	9.5	8.9	8.5	8.4
35	9.9	9.1	7.9	7.0	6.6	6.4
50	9.6	8.8	7.6	7.0	6.9	6.9

vocabulary that serves as the basis for training and therefore, it is not very effective in cases an extension of the vocabulary is necessary.

3.3. Machine Learning

Latest advances in machine learning area for learning sequences with high dependencies among their elements (Bahdanau et al., 2014) has raised community interest for their applications in handwriting synthesis. Two interesting examples in this area are the works of Ahmad and Fink (2015) and Graves (2013).

In the work "Training an Arabic handwriting recognizer without a handwritten training data set" Ahmad and Fink (2015) developed a classification system based on font typefaces and HMMs. What makes this work interesting in terms of machine learning is that it employs an unsupervised HMM adaptation. The basic idea here is to find the optimal value of HMM parameter, θ, using Maximum Likelihood Linear Regression (MLLR) (Saleem et al., 2009). The MLLR was used together with the adjustment of HMM parameters in the iterative training process, where the worst results at each iteration are excluded, resulting in an improvement of the final accuracy. For the experiments, 8 font typefaces (Tahoma, Rekka, Diwani, Arabic Typesetting, Traditional Arabic, Thuluth, Zarnew and Naskh) were adopted and the dataset IFN/ENIT (Pechwitz et al., 2002) was used for system evaluation. The results for each typeface are shown in Table 3. Together with each typeface results some other experimetns were conducted using all typefaces, unsupervised adaptation (Young et al., 2002), and multi-stream HMMs (Ahmad et al., 2014) as can be seen in Table 4.

Table 3. Ahmad e Fink's results for each typeface.

Typeface	Word Recognition (%)
Tahoma	04.31
Rekaa	07.28
Diwani	10.01
Arabic Typesetting	11.25
Traditional Arabic	12.87
Thuluth	17.67
Zarnew	18.75
Naskh	26.92

As can be seen, the use of typefaces for handwriting synthesis to generate a classifier

Document Processing Company, Ex. 2035
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 165 of 404

Table 4. Results Ahmad e Fink

System	Samples			
	d	e	f	s
Typeface font (Naskh)	26.92	22.10	24.10	27.39
All typefaces together	61.35	55.84	55.14	51.94
Text images from all typefaces together + unsupervised adaptation	70.47	66.53	60.93	54.74
Text images from all typefaces together + unsupervised adaptation + Multi-stream HMMs	91.61	89.61	86.58	73.11

for a real scenario results very interestingly mainly because the facility of generating this dataset.

However, it is required a more careful analysis about how a system performance would behave in case of using the Latin or Cyrillic alphabet, where the differences between handwritten and printed alphabets are more considerable.

3.4. Handwriting Synthesis with Recurrent Network

Graves in "Generating sequences with recurrent neural networks" (Graves, 2013) carries out an assessment of the power of recurrent neural network systems with Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) memory cells in the generation of sequences. The study evaluates how the system output (named prediction networks) can be used to generate text with the *Penn Treebank dataset* (Marcus et al., 1993). Moreover, by using a more complex example with the *Wikipedia* dataset (Hutter, 2012), it shows the power of the system to learn information like URL formats and how HTLM tags works. Finally, a synthesis model based on the IAMOnDB (Liwicki and Bunke, 2005b) is developed.

The network proposed for the case of synthesis consists of an input layer with 3 units (position x, y and end of stroke information), 3 hidden layers of 400 LSTM neurons and an output layer with 121 outputs feed into a mixture density function. This function generates the 3 output values: position x, y and end of stroke information. Between each hidden layer there is a character-level layer being responsible for making the connection between the following points and the text itself. For each entry at time t, the network's objective is to predict which will be the next entry at $t + 1$ in the sequence.

Once the network has been trained, it is presented with a text and an entry $[0, 0, 0]$, and it is expected that at each step its output predicts the next value of pen movement. The results are very promising and show that synthesized handwritten text may well be confused with one produced by real people. However, any assessment of this fact has been held yet, and any check has been made to see whether the approach may or may not improve existing foundations. Nevertheless, in our experiments, we have observed a very interesting point of this approach when we analyze synthesized handwriting outputs of the network trained with the English dataset IAMOnDB. Such outputs correspond to some Portuguese words (Figure 1) and sentences (Figure 2), which were presented to the network. This somehow shows the ability of the network to synthesized handwritten text in a specific language

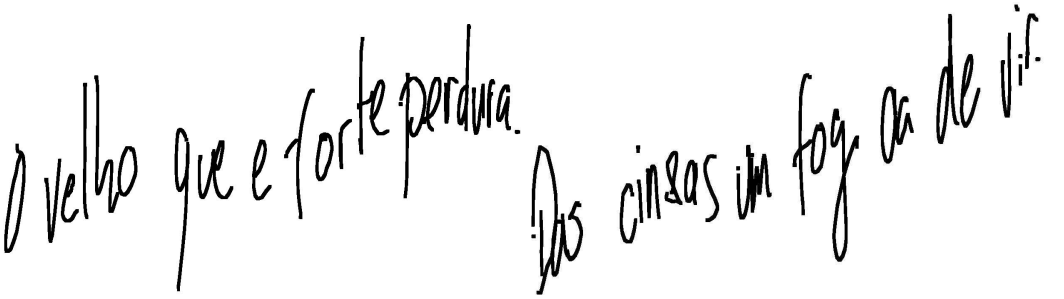
Licensed to Cerence Operating Company - Exhibit 1002
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 166 of 404

(Portuguese) using an already know alphabet learned from a different language (English).



(a) Brasil (b) Setembro (c) Abril

Figure 1. Samples of Portuguese words generated by the network.



(a) O velho que é forte perdura (b) Das cinzas um fogo a de vir

Figure 2. Sample of Portuguese phrases generated by the network.

4. Printed-Text Dataset Generation

For the printed text dataset generation we use printed scanned documents (e.g magazines, contracts, forms and journals) to compile a new dataset. The images are segmented and different Optical Character Recognition (OCR) tools are used to label them. The recognized image text are combined by different heuristics to define the class label. The proposed system is divided into three main parts: pre-processing, processing and decision. The Figure 3 presents the basic structure of the printed text dataset generation system.

The first step is the pre-processing, which the input images are prepared for processing. A variety of operations can be applied at this step such as format conversion, rescaling, binarization and noise removal. Right after the pre-processing, a variety of OCR systems is used to extract text on the processed images. Finally, having the image and many OCR outputs, the last step is responsible for building the final word dataset using tiebreaker heuristics.

4.1. Pre-Processing

In this step, the input images are prepared for processing. The main goal of the pre-processing is to enhance the visual of the images (by removing or reducing noise, for in-

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 167 of 404

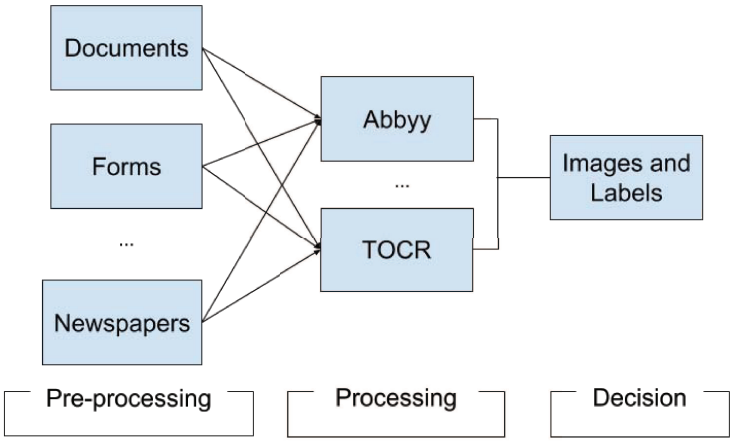


Figure 3. Structure of the printed-text dataset generation system.

stance) and improve the manipulation of the datasets. Some OCR systems, for example, only deal with TIFF format, others only accept or work better with binarized images.

It's very important to point out that this step modifies the image and the created dataset has to be formed by the original image. For this reason, after the pre-processing step, there will be two versions of each image: the original, without any modifications and the pre-processed, which is the enhanced image.

Most OCR systems automatically do various image processing operations, but depending on the input images, there will inevitably be cases where the results aren't good, directly impacting the quality of the system. The most common image processing operations are:

Format conversion. Some OCR systems only accept specific image formats. Transym OCR [3], for instance, can read bitmap (.bpm) and TIFF (.tiff) images. If you wish to process any other formats, you need to convert them to the input formats. This is what the format conversion operation is responsible for.

Rescaling. Rescaling refers to changing the size of an image in terms of both resolution and dots per inch (dpi). Regarding text size, for example, Tesseract suggests that accuracy drops off bellow $10pt \times 300dpi$. Rescaling is also important when the input images are on a higher scale than necessary and you want to speed up the processing. Therefore, to achieve better results, you probably need to rescale the input images. To ensure the image quality we suggest the use of algorithms based on bicubic Interpolation (Bourke, 2001).

Thresholding. or binarization basically means converting to a black and white image. Most OCR systems automatically binarize the images, but the results can be suboptimal and they also recommend that the used manually binarize the image before processing. This is the case of ABBYY FineReader, TOCR and Tesseract, for example. The thresholding algorithms can be applied globally (like Otsu, Pun, Kapur, Renyi, two peaks, and percentage

[3] http://www.transym.com/

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 168 of 404

of black) and locally (Niblack, Sauvola, White and Bernsen) algorithms. For a general view of all these methods and much more, we suggest the reading of Sezgin and Sankur survey (Sezgin et al., 2004). Although in average local binarization methods perform slightly better than the global ones, there is a large performance variance. In some cases, some global methods have a very good performance and some local ones are close to the worst options (Stathis et al., 2008). The classical Sauvola algorithm is a very stable method for a general propose document (Sauvola and Pietikäinen, 2000).

Noise removal. Noise usually originates from the image acquisition process and often results in unrealistic intensity of pixels that can make the text of an image more difficult to read. There are several techniques for this purpose. Low-pass, high-pass, band-pass spatial filtering, mean filtering and median filtering are few examples. For a general view of all these methods, we suggest this survey (Chakraborty and Blumenstein, 2016).

4.2. Processing

After the preprocessing, there are two versions of each image: the original and the processed. In the processing step, the processed images are presented for each the OCR system in order to recognize the words on the image.

This step is also responsible for parsing the OCR engines' output. If you plan to test different settings you can optionally store the output in a structured text format (for example, XML) containing, for each image, the image path and the OCR output. Storing the OCR output is important to save time by avoiding rerunning the engines. The output usually contains the word or character position, text orientation, the text itself and the confidence. It's also important to store this information for the next step.

In order to speed up this step, you can optionally run the OCR systems in parallel. Note that depending on the product, there are licensing restrictions and this might not be possible (i.e. there are licenses that only allow you to run one OCR engine on one CPU core at a time). To remove these restrictions you'll probably need to purchase extra licenses.

4.3. Decision

Using the OCR output from the last step, the Decision is responsible for generating the dataset containing images of words and its respective label. These heuristics are mapped in two operators: the Matching Operator and the Confidence Operator.

Confidence Operator. Most of OCR engines provide the confidence per character (a number between 0 and 1). The purpose of this operator is to evaluate the word confidence based on each character confidence. This value can be calculated using median, average and minimum of the character confidences. We suggest the use of the minimum score of each word as the confidence.

Matching Operator. The matching operator decides which word will be used in the final generated dataset. For example, let's say both Engine A and Engine B recognize the word

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 169 of 404

"house" of a given image, but the Engine C doesn't. The matching operator decides whether this word will be selected for the final dataset or not.

The Confidence Operator and the Matching Operator work together for the final image selection. While the former evaluates the word confidence, the latter uses a threshold that can ignore words with confidence less than this threshold.

At the same time, the Matching Operator uses few more heuristics for its decision. They were called Match All and Keep All. On the Keep All heuristic, the final dataset will contain all non-ignored words from all engines, only respecting the threshold. For example, suppose the Engine A recognized the words "victory" and "house" in one image and the Engine B recognized the words "victory" and "ouse" for the same image, the words "victory", "house" and "ouse" will form the final dataset regardless of the different results. When using the Keep All, the final dataset has more images than the Match All heuristic. This heuristic is also important to understand each OCR engine behavior.

Furthermore, the Match All heuristic will only use the non-ignored words that were recognized by all engines. On the previous example, the final dataset will only contain the word "victory", because that was the only non-ignore word recognized by both engines.

Conclusion

We present a analysis of the current state of datasets, showing how they are already built and they main features. As is pointed by us those datasets suffer from the problem of the amount of data and in some cases diversity of data or the versions of data (a lot of samples to a sentence or word). To overcome this problem we present two paths to be followed, the first one mostly focused in handwriting based on synthesis with a review of the current state-of-the art of the techniques and then showing how to use then to extend beyond the limitation of diversity and variety. The second path is focused in Optical Character Recognition and focus in using a combination of systems to produce a better solution based on a tiebreak decision using confidence and matching operators. With those two paths, we believe that we can overcome the current problem of the amount of data and improve the current datasets and consequently the recognition models.

Acknowledgment

The authors would like to thank the CNPQ for supporting the development of this chapter through the research projects granted by "Edital Universal" (Process 444745/2014-9) and "Bolsa de Produtividade DT" (Process 311912/2015-0). In addition, the authors also acknowledge the Document Solutions for providing several real image samples useful for the development of this research.

References

Ahmad, I. and Fink, G. A. (2015). Training an arabic handwriting recognizer without a handwritten training data set. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 476–480. IEEE.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 170 of 404

Ahmad, I., Fink, G. A., and Mahmoud, S. A. (2014). Improvements in sub-character hmm model based arabic text recognition. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 537–542. IEEE.

Al-Muhtaseb, H., Elarian, Y., and Ghouti, L. (2011). Arabic handwriting synthesis. In *First International Workshop on Frontiers in Arabic Handwriting Recognition, 2010*.

Antonacopoulos, A., Bridson, D., Papadopoulos, C., and Pletschacher, S. (2009). A realistic dataset for performance evaluation of document layout analysis. In *2009 10th International Conference on Document Analysis and Recognition*, pages 296–300.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bourke, P. (2001). Bicubic interpolation for image scaling.

Brunessaux, S., Giroux, P., Grilhres, B., Manta, M., Bodin, M., Choukri, K., Galibert, O., and Kahn, J. (2014). The maurdor project: Improving automatic processing of digital documents. In *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*, pages 349–354.

Chakraborty, A. and Blumenstein, M. (2016). Marginal noise reduction in historical handwritten documents–a survey. In *Document Analysis Systems (DAS), 2016 12th IAPR Workshop on*, pages 323–328. IEEE.

Chen, H.-I., Lin, T.-J., Jian, X.-F., Shen, I., Chen, B.-Y., et al. (2015). Data-driven handwriting synthesis in a conjoined manner. In *Computer Graphics Forum*, volume 34, pages 235–244. Wiley Online Library.

Diem, M., Kleber, F., and Sablatnig, R. (2011). Text classification and document layout analysis of paper fragments. In *2011 International Conference on Document Analysis and Recognition*, pages 854–858.

Dinges, L., Al-Hamadi, A., Elzobi, M., El-etriby, S., and Ghoneim, A. (2015). Asm based synthesis of handwritten arabic text pages. *The Scientific World Journal*, 2015.

Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

Elarian, Y., Ahmad, I., Awaida, S., Al-Khatib, W. G., and Zidouri, A. (2015). An arabic handwriting synthesis system. *Pattern Recognition*, 48(3):849–861.

Fischer, A., Frinken, V., Bunke, H., and Suen, C. Y. (2013). Improving hmm-based keyword spotting with character language models. In *2013 12th International Conference on Document Analysis and Recognition*, pages 506–510. IEEE.

Fischer, A., Frinken, V., Fornés, A., and Bunke, H. (2011a). Transcription alignment of latin manuscripts using hidden markov models. In *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, pages 29–36. ACM.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 171 of 404

Fischer, A., Indermühle, E., Bunke, H., Viehhauser, G., and Stolz, M. (2010a). Ground truth creation for handwriting recognition in historical documents. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, DAS '10, pages 3–10, New York, NY, USA. ACM.

Fischer, A., Indermuhle, E., Frinken, V., and Bunke, H. (2011b). Hmm-based alignment of inaccurate transcriptions for historical documents. In *2011 International Conference on Document Analysis and Recognition*, pages 53–57.

Fischer, A., Keller, A., Frinken, V., and Bunke, H. (2012). Lexicon-free handwritten word spotting using character hmms. *Pattern Recognition Letters*, 33(7):934–942.

Fischer, A., Riesen, K., and Bunke, H. (2010b). Graph similarity features for hmm-based handwriting recognition in historical documents. In *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, pages 253–258.

Fischer, A., Wuthrich, M., Liwicki, M., Frinken, V., Bunke, H., Viehhauser, G., and Stolz, M. (2009). Automatic transcription of handwritten medieval documents. In *Virtual Systems and Multimedia, 2009. VSMM'09. 15th International Conference on*, pages 137–142. IEEE.

Frinken, V., Fischer, A., Manmatha, R., and Bunke, H. (2012). A novel word spotting method based on recurrent neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2):211–224.

Garris, M. D., Blue, J. L., Candela, G. T., et al. (1997). Nist form-based handprint recognition system (release 2.0).

Gatos, B., Louloudis, G., Causer, T., Grint, K., Romero, V., Snchez, J. A., Toselli, A. H., and Vidal, E. (2014). Ground-truth production in the transcriptorium project. In *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*, pages 237–241.

Graves, A. (2012). Offline arabic handwriting recognition with multidimensional recurrent neural networks. In *Guide to OCR for Arabic Scripts*, pages 297–313. Springer.

Graves, A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Guyon, I. (1996). Handwriting synthesis from handwritten glyphs. In *Proceedings of the Fifth International Workshop on Frontiers of Handwriting Recognition*, pages 140–153. Citeseer.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hutter, M. (2012). The human knowledge compression contest. 2012. *U RL http://prize. hutter1. net*.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 172 of 404

Indermühle, E., Liwicki, M., and Bunke, H. (2010). Iamondo-database: an online handwritten document database with non-uniform contents. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pages 97–104. ACM.

Jawahar, C. and Balasubramanian, A. (2006). Synthesis of online handwriting in indian languages. In *Tenth International Workshop on Frontiers in Handwriting Recognition*. Suvisoft.

Kermorvant, C. and Louradour, J. (2010). Handwritten mail classification experiments with the rimes database. In *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, pages 241–246.

Kleber, F., Fiel, S., Diem, M., and Sablatnig, R. (2013). Cvl-database: An off-line database for writer retrieval, writer identification and word spotting. In *2013 12th International Conference on Document Analysis and Recognition*, pages 560–564. IEEE.

Lazzara, G. and Géraud, T. (2014). Efficient multiscale sauvola's binarization. *International Journal on Document Analysis and Recognition (IJDAR)*, 17(2):105–123.

LeCun, Y., Cortes, C., and Burges, C. J. (2012). The mnist database of handwritten digits, 1998. *Available electronically at http://yann. lecun. com/exdb/mnist*.

Liwicki, M. and Bunke, H. (2005a). Handwriting recognition of whiteboard notes. In *Proc. 12th Conf. of the Int. Graphonomics Society*, pages 118–122.

Liwicki, M. and Bunke, H. (2005b). Iam-ondb-an on-line english sentence database acquired from handwritten text on a whiteboard. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pages 956–961. IEEE.

Liwicki, M. and Bunke, H. (2006). Hmm-based on-line recognition of handwritten whiteboard notes. In *Tenth international workshop on frontiers in handwriting recognition*. Suvisoft.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Marti, U.-V. and Bunke, H. (1999). A full english sentence database for off-line handwriting recognition. In *Document Analysis and Recognition, 1999. ICDAR'99. Proceedings of the Fifth International Conference on*, pages 705–708. IEEE.

Martín-Albo, D., Plamondon, R., and Vidal, E. (2014). Training of on-line handwriting text recognizers with synthetic text generated using the kinematic theory of rapid human movements. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 543–548. IEEE.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 173 of 404

Menasri, F., Louradour, J., Bianne-Bernard, A.-L., and Kermorvant, C. (2012). The a2ia french handwriting recognition system at the rimes-icdar2011 competition. In *IS&T/SPIE Electronic Imaging*, pages 82970Y–82970Y. International Society for Optics and Photonics.

Padmanabhan, R. K., Jandhyala, R. C., Krishnamoorthy, M., Nagy, G., Seth, S., and Silversmith, W. (2009). Interactive conversion of web tables. In *International Workshop on Graphics Recognition*, pages 25–36. Springer.

Pechwitz, M., Maddouri, S. S., Märgner, V., Ellouze, N., Amiri, H., et al. (2002). Ifn/enit-database of handwritten arabic words. In *Proc. of CIFED*, volume 2, pages 127–136. Citeseer.

Pham, V., Bluche, T., Kermorvant, C., and Louradour, J. (2014). Dropout improves recurrent neural networks for handwriting recognition. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 285–290. IEEE.

Plamondon, R., O'Reilly, C., Galbally, J., Almaksour, A., and Anquetil, É. (2014). Recent developments in the study of rapid human movements with the kinematic theory: Applications to handwriting and signature synthesis. *Pattern Recognition Letters*, 35:225–235.

Rath, T. M. and Manmatha, R. (2007a). Word spotting for historical documents. *International Journal of Document Analysis and Recognition (IJDAR)*, 9(2):139–152.

Rath, T. M. and Manmatha, R. (2007b). Word spotting for historical documents. *International Journal of Document Analysis and Recognition (IJDAR)*, 9(2-4):139–152.

Saleem, S., Cao, H., Subramanian, K., Kamali, M., Prasad, R., and Natarajan, P. (2009). Improvements in bbn's hmm-based offline arabic handwriting recognition system. In *2009 10th International Conference on Document Analysis and Recognition*, pages 773–777. IEEE.

Sauvola, J. and Pietikäinen, M. (2000). Adaptive document image binarization. *Pattern recognition*, 33(2):225–236.

Schlapbach, A., Liwicki, M., and Bunke, H. (2008). A writer identification system for on-line whiteboard data. *Pattern Recogn.*, 41(7):2381–2397.

Sezgin, M. et al. (2004). Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic imaging*, 13(1):146–168.

Shahab, A., Shafait, F., Kieninger, T., and Dengel, A. (2010). An open approach towards the benchmarking of table structure recognition systems. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pages 113–120. ACM.

Stathis, P., Kavallieratou, E., and Papamarkos, N. (2008). An evaluation technique for binarization algorithms. *J. UCS*, 14(18):3011–3030.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 174 of 404

Toselli, A. H. and Vidal, E. (2015). Handwritten text recognition results on the bentham collection with improved classical n-gram-hmm methods. In *Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing*, pages 15–22. ACM.

Yalniz, I. Z. and Manmatha, R. (2011). A fast alignment scheme for automatic ocr evaluation of books. In *2011 International Conference on Document Analysis and Recognition*, pages 754–758. IEEE.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al. (2002). The htk book. *Cambridge University engineering department*, 3:175.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 175 of 404

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

PART II.
ANALYSIS AND APPLICATIONS

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

In: Handwriting: Recognition, Development and Analysis ISBN: 978-1-53611-937-4
Editors: Byron L. D. Bezerra et al. © 2017 Nova Science Publishers, Inc.

Chapter 7

MATHEMATICAL EXPRESSION RECOGNITION

***Francisco Álvaro*[1,*], *Joan Andreu Sánchez*[2,†] and *José Miguel Benedí*[2,‡]**
[1]WIRIS Math
[2]Pattern Recognition and Human Language Technologies Research Center,
Universitat Politècnica de València, València, Spain

1. Introduction

Mathematical notation is a well-known language that has been used all over the world for hundreds of years. Despite the great number of cultures, languages and even different writing scripts, mathematical expressions constitute a universal language in many fields. During the last century and in particular since the development of the Internet, digital information represents the best resource for accessing and sharing data. Therefore, it is necessary to digitize documents and to input mathematical expressions directly into computers.

Although most people know how to read or write mathematical expressions, introducing them into a computer device usually requires learning specific notations or knowledge of how to use a certain editor. Mathematical expression recognition intends to fill this gap between the knowledge of a person and the language computers understand.

1.1. Problem Description

Mathematical expression recognition is a classical problem of pattern recognition and its goal is to obtain the mathematical expression encoded in a given input sample. In this field we distinguish different types of mathematical expressions that require specific treatment. In this section, we will first describe the taxonomy of mathematical expressions considered in this problem. Afterwards, we state the main tasks that a math recognition system has to deal with.

First, a mathematical expression can be either printed or handwritten. Printed formulae are commonly easier to recognize than handwritten expressions because they tend to be

*E-mail address: falvaro@wiris.com.
†E-mail address: jandreu@prhlt.upv.es.
‡E-mail address: jmbenedi@prhlt.upv.es.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

more regular. Thus, individual elements and the relation between them can be determined more consistently. Handwriting introduces more variability in the shape of the symbols and the relationship between them. Also, there are many different writers and writing styles, thus, handwritten mathematical expression recognition is more challenging. Figures 1 and 2 show an example of the printed and handwritten version of the same formula.

$$a_0 = x^2 + \frac{\sqrt{\pi}}{3}$$

Figure 1. Example of printed mathematical expression.

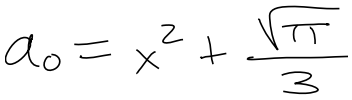


Figure 2. Example of handwritten mathematical expression.

Regarding the input representation, we consider the problem to be *off-line* if the mathematical expression is represented as an image, i.e. a matrix of pixels. On the other hand, a mathematical expression is considered to be *on-line* when it has been acquired using a device which provides us with the temporal information of the writing, i.e. the input is a time sequence of points.

The representation of mathematical expressions is based on different primitives depending on the type of expression. Off-line expressions are usually based on connected components.

Definition 1. *A* connected component *of an image is a set of adjacent foreground pixels, where pairs of pixels are connected in such a way that they are neighbors in an 8-connected sense.*

The primitives for representing on-line mathematical expressions are usually strokes.

Definition 2. *A* stroke *is the sequence of points drawn from when a pen touches the surface until the user lifts the pen from the surface.*

These definitions of primitives can be seen in the examples of Figures 1 and 2. In the printed expression of Figure 1, symbols π and $+$ are made up of one connected component, and symbol $=$ is composed of two connected components. If the handwritten expression of Figure 2 was on-line, symbol π would be composed of three strokes, symbol $+$ would be composed of two strokes and number 0 would be drawn using just one stroke. But if the handwritten expression of Figure 2 was off-line, the instance of symbol π would be composed of two connected components, and the instances of symbol $+$ and number 0 would be made up of one connected component.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 180 of 404

As a result of these differences, the problem of recognizing mathematical expressions has three possible scenarios: off-line printed math expression recognition, and both on-line and off-line handwritten math expression recognition.

Automatic recognition of mathematical notation is traditionally divided into three problems (Chan and Yeung, 2000; Zanibbi and Blostein, 2012): symbol segmentation, symbol recognition and structural analysis. The issues related to each of the problems mentioned above are detailed below.

1.1.1. Symbol Segmentation

Symbol segmentation is the problem of determining what parts of the input expression form a mathematical symbol. Depending on the type of expression, on-line or off-line, there are different issues to cope with.

Segmentation of off-line mathematical expressions is usually based on computing the connected components of the image.



(a) (b) (c) (d)

Figure 3. Symbol segmentation problems in off-line mathematical expression recognition based on connected components.

Figure 3 shows examples of the problems related to segmentation of off-line mathematical symbols. First, many symbols are made up of more than one connected component by definition and have to be grouped (Figure 3-a). A difficult problem in off-line segmentation is when the connected components of two different symbols are touching and have to be split (Figure 3-b). Also, some symbols can be broken into several components due to image degradation (Figure 3-c). Finally, images can contain noise that produces additional components that do not form any symbol and should be grouped or ignored (Figure 3-d).

Segmentation of on-line expressions is commonly based on strokes, although symbol segmentation could also be based on connected components where it would have the same problems than in off-line segmentation. Stroke-based segmentation can handle the problem of touching symbols provided that two touching symbols are written in different strokes. An equivalent problem would appear if two symbols share a single stroke.

Mathematical symbols can be made up of one or more strokes. For instance, in off-line segmentation a plus sign ($+$) is detected as a connected component, but in on-line segmentation it has two strokes that must be merged. Figure 2 shows several multi-stroke symbols whose strokes must be grouped: π, $+$ and $=$. Finally, although it is less common than noise in images, small strokes can be introduced by the user that do not belong to any symbol of the mathematical expression and should be discarded.

Context information is important to determine the segmentation of an input expression. In Figure 4-a we can see several segmentation hypotheses that are reasonable if context is not taken into account. For instance, the plus-minus sign ($\pm$) could be split into a plus sign and a horizontal line, or the fraction line and the top stroke of the number five could be merged as an equals sign (Awal et al., 2009). Figure 4-b shows ambiguities due to

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 181 of 404

handwriting production, in that the expression could have several valid interpretations with alternative segmentations: $\boxed{1-1<x}$, $\boxed{1-kx}$ or $\boxed{H<x}$.

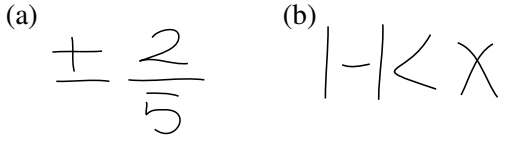


Figure 4. Handwritten mathematical expressions showing several examples of ambiguities in symbol segmentation.

1.1.2. Symbol Recognition

Mathematical symbol recognition aims to identify the symbol encoded by a given hypothesis. Commonly, in off-line recognition a hypothesis is an image, and in on-line recognition a symbol hypothesis is a set of strokes. There are many sets of symbols in mathematical notation: the Latin alphabet ($a-z$, $A-Z$), the Greek alphabet (α, β, γ, π, $\sum$, ...), numbers ($0-9$), operators ($+$, $-$, $/$, $\int$, $\sqrt{}$, ...) and more (∞, $\rightarrow$, $\forall$, $\{$, $\}$, ...).

Some of the symbols in mathematics are very similar, and there are even symbols that are represented by the same shape, for instance, the number 0 and the letter o, or the letter x and the operator $\times$. Context information of the mathematical expression can help to solve the ambiguities in symbol recognition and determine the correct symbol class. We can see an example in Figure 5 where the same shape represents a letter (x) in the upper expression and an Cartesian product operator ($\times$) in the lower expression.

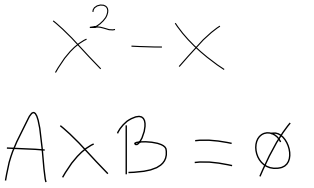


Figure 5. Example of symbol classification depending on the context in the mathematical expression. The same symbol shape is classified as a letter in top expression ($x^2 - x$) and as a product operator between sets in bottom expression ($A \times B = \emptyset$).

1.1.3. Structural Analysis

A mathematical expression is made up of symbols and of the different relationships between them. The final objective of mathematical expressions recognition will not only be recognition of the symbols that make up the mathematical expression, but it must include the recognition of the structure that relates them.

The structural analysis of equations requires determining two-dimensional relationships between symbols or sets of symbols. In Figures 1 and 2 we can see examples of the most

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 182 of 404

common relations: *subscript* between symbols a and 0; *superscript* between symbols x and 2; *below* between the elements of the fraction; *inside* in the square root; and the *right* relationship for the horizontal concatenation of the elements in the expression. There are other relations like radicals ($\sqrt[x]{\ }$), left scripts (${}_b^a x$) or the complex structure of matrices.

Relations between symbols can be ambiguous in several situations, meaning that detecting the correct relationship might require knowing the symbols or reliance on language models. Figure 6 shows an example where the relationship between two symbols in a mathematical expression requires taking into account the entire expression.

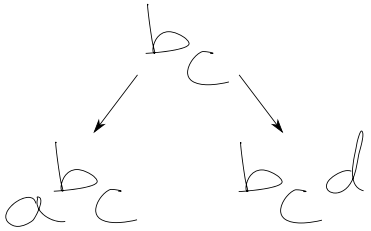


Figure 6. Example of subscript and superscript relationships that cannot be determined locally (Chan and Yeung, 2000).

Normally, the structure of a mathematical expression is represented as a tree. Figure 7 shows the same expression encoded by three types of trees. In relational trees, the mathematical symbols are the leaves of the tree and the internal nodes represent the relationship, while in symbol layout trees each node is a symbol and the edges indicate the relationships. In operator trees, leaves represent symbols and each node contains the operation that computes the expression bottom-up.

Finally, structural analysis is important to determine the correct segmentation and the identity of the recognized symbols. Some symbols can only be correctly classified if their spatial relationships are taken into account. For example, a horizontal line can represent a minus operator, a fraction bar or can be part of a symbol (e.g. $=$, $\leq$ or $\pm$). Context information is crucial in solving these ambiguities, as we can see in the example of Figure 4.

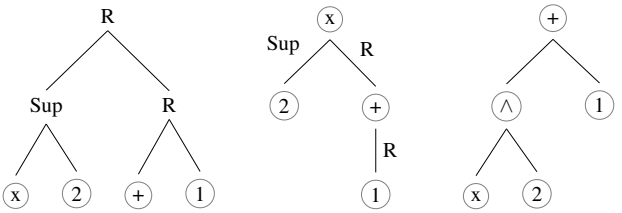


Figure 7. Example of tree representation of the expression $x^2 + 1$. Left to right: relational tree, symbol layout tree, and operator tree.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 183 of 404

2. State of the Art

The problem of automatic mathematical expression recognition has been studied for decades (Anderson, 1967). Many approaches have been proposed (Chou, 1989; Zanibbi et al., 2002; MacLean and Labahn, 2013) and we could group the proposals into three big families: projection-based methods, graph-based methods and grammar-based methods. Some of the solutions proposed in this field are reviewed in the following section.

2.1. Projection-Based Approaches

Mathematical expressions can be seen as nested structures of symbols. For instance, a mathematical expression containing a fraction has two expressions: numerator and denominator. Some proposals in the literature are based on recursively dividing the mathematical formula into sub-expressions by means of projection profiles (Okamoto and B., 1991), the X-Y cut algorithm (Ha et al., 1995) or using prior knowledge about the structure of mathematical notation (Faure and Wang, 1990).

These types of method are top-down processes where the mathematical zone is commonly divided by left-to-right vertical division, then each sub-zone is divided by top-to-bottom horizontal division, and this process is repeated until primitive objects are reached. An example of recursive decomposition is shown in Figure 8.

These approaches effectively decompose the mathematical expression into smaller subproblems and tend to be fast. However, some cases require special treatment since some sub-expressions overlap (e.g. square roots). Also, sloped expressions can be challenging for this methodology because projections might not clearly divide the sub-expressions.

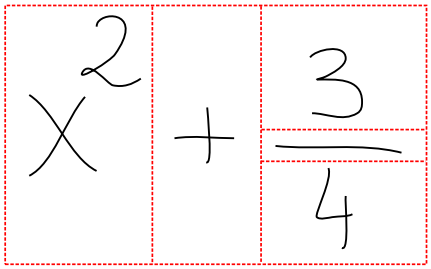


Figure 8. Abstraction of projection-based approaches for math expression recognition. Regions are recursively divided: left-to-right, top-bottom.

2.2. Graph-Based Approaches

Another group of methods are based on graphs or trees. These formalisms represent a proper structure to deal with the 2D spatial relationships and the structure representation of a mathematical expression, and also there are many efficient algorithms for them. Figure 9 shows an example of a graph built from the primitives of a mathematical expression. The edges in the graph can be weighted according to different criteria and the recognized ex-

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 184 of 404

pression can be obtained by computing the minimum spanning tree. Many approaches of this group have been proposed and we briefly summarize some of them below.

Eto and Suzuki (2001) developed a model for printed math expression recognition that computed the minimum spanning tree of a network representation of the expression. Tapia and Rojas (2004) presented a proposal for online recognition also based on constructing the minimum spanning tree and using symbol dominance. Zanibbi et al. (2002) recognized an expression as a tree, and proposed a system based on a sequence of tree transformations. Lehmberg et al. (1996) defined a net so that the sequence of symbols within the handwritten expression was represented by a path through the graph. Shi et al. (2007) presented a similar system where symbol segmentation and recognition were tackled simultaneously based on graphs. They then generated several symbol candidates for the best segmentation, and the recognized expression was computed in the final structural analysis (Shi and Soong, 2008).

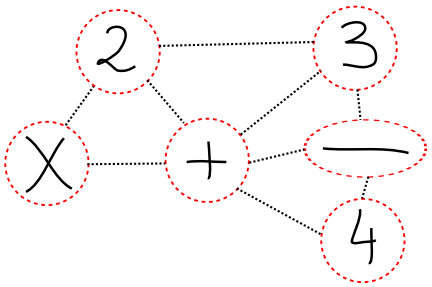


Figure 9. Abstraction of graph-based approaches for mathematical expression recognition. Primitives are connected to create a graph used for the computation of the recognized expression.

This group of approaches generally results in efficient algorithms for recognizing formulas, and trees and graphs are proper models for representing mathematical expressions. However, context-free dependencies are not naturally modeled in most of these structures. Also, some approaches require a one-dimensional order, but mathematical notation is 2D. Therefore, the order is often achieved by detecting baselines and exploiting the left-to-right reading order. But errors in the baseline detection cannot be solved in further steps. Another option to obtain a one-dimensional order in online recognition is to assume that symbols are written with consecutive strokes, which limits the set of accepted inputs.

2.3. Grammar-Based Approaches

Given the well-defined structure of mathematical notation, many approaches are based on grammars because they constitute a natural way to model this problem. In fact, the first proposals on mathematical expression recognition were grammar-based (Anderson, 1967; Chou, 1989). Since then, different studies have been developed using different types of grammars. For instance, Chan and Yeung (2001) used definite clause grammars, the model of Lavirotte and Pottier (1998) was based on graph grammars, Yamamoto et al. (2006) presented a system using Probabilistic Context-Free Grammars (PCFG), and MacLean and

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 185 of 404

Labahn (2013) developed an approach using relational grammars and fuzzy sets. Despite previous approaches using different types of grammars, the methodology is based on the same process.

Grammars allow us to model complex structural relationships by means of the rules, which combine sub-problems to construct larger hypotheses (see Figure 10). In this chapter we will focus on solutions based on PCFG since we will detail an approach based on this formalism in the next section.

Proposals based on PCFG use grammars to model the structure of the expression, but the recognition systems are different. Garain and Chaudhuri (2004) proposed a system that combines online and offline information in the structural analysis. First, they created online hypotheses based on determining baselines in the input expression, and then offline hypotheses using recursive horizontal and vertical splits. Finally they used a context-free grammar to guide the process of merging the hypotheses. Yamamoto et al. (2006) presented a version of the CYK algorithm for parsing 2D-PCFGs with the restriction that symbols and relations must follow the writing order. They defined probability functions based on a region representation called "hidden writing area". Pra and Hlav (2007) described a system for offline recognition using 2D context-free grammars. Their proposal was penalty-based so that weights were associated with regions and syntactic rules. The model proposed by Awal et al. (2014) considers several segmentation hypotheses based on spatial information, and the symbol classifier has a rejection class in order to avoid incorrect segmentations. Álvaro et al. (2016) developed an integrated model based on parsing 2D-PCFG where the recognition process globally optimizes the most likely expression according to several probabilistic sources. In the following section, we will further detail this proposal as an example of a solution for math expression recognition.

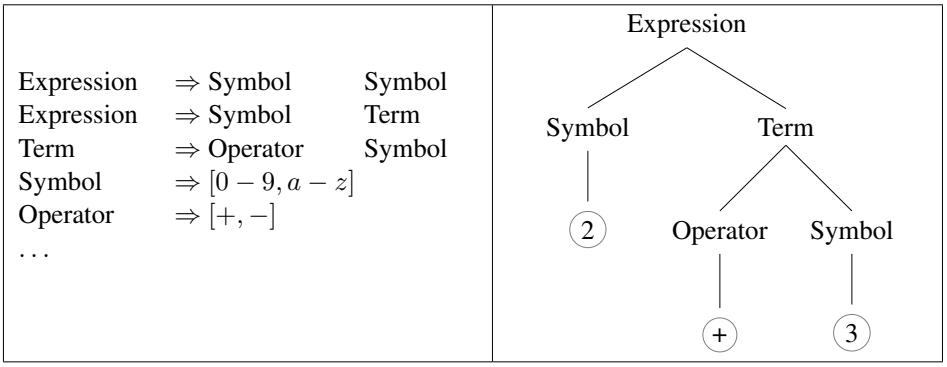


Figure 10. Abstraction of grammar-based approaches for mathematical expression recognition.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 186 of 404

3. Integrated Grammar-Based Proposal for Mathematical Expression Recognition

In on-line handwritten mathematical expression recognition, the input is a sequence of strokes, and these strokes are in themselves a sequence of points. Figure 11 shows an example of the input for a mathematical expression. As you can see, the temporal sequence

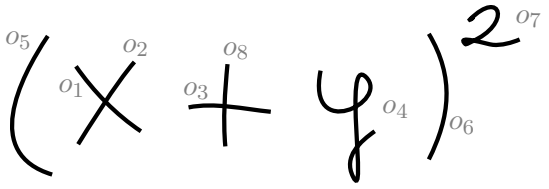


Figure 11. Example of input for an on-line handwritten math expression. The order of the input sequence of strokes is labeled ($\boldsymbol{o} = o_1 o_2 \ldots o_8$).

of strokes does not correspond necessarily to the sequence of symbols that it represents. For example, we can see that the user first wrote the sub-expression $x - y$, then the user added the parentheses and its superscript $(x - y)^2$, finally converting the subtraction into an addition $(x + y)^2$. This example shows that some symbols might not be made up of consecutive strokes (e.g. the $+$ symbol in Figure 11). This means that the mathematical expression would not be correctly recognized if it was parsed monotonically with the input, i.e. processing the strokes in the order in which they were written. Meanwhile, the sequence of symbols that make up a sub-expression does not have to respect the writing order (e.g. the parentheses and the sub-expression they contain in Figure 11).

Given a sequence of input strokes, the output of a mathematical expression recognizer is usually a sequence of symbols (Shi et al., 2007). However, we consider that a significant element of the output is the structure that defines the relationship between the symbols which make up the final mathematical expression. As mentioned above, we propose modeling the structural relationships of a mathematical expression using a statistical grammatical model. By doing so, we define the problem of mathematical expression recognition as obtaining the most likely parse tree given a sequence of strokes. Figure 12 shows a possible parse tree for the expression given in Figure 11, where we can observe that a (context-free) structural model would be appropriate due to, for instance, structural dependencies in bracketed expressions. The output parse tree represents the structure that relates all the symbols and sub-expressions that make up the input expression. The parse tree derivation produces the sequence of pre-terminals that represent the recognized mathematical symbols. Furthermore, to generate this sequence of pre-terminals, we must take into account all stroke combinations in order to form the possible mathematical symbols.

Taking these considerations into account, two main problems have been observed. First, the segmentation and recognition of symbols is closely related to the alignment of mathematical symbols to strokes. Second, the structural analysis of a mathematical expression addresses the problem of finding the parse tree that best accounts for the relationships between different mathematical symbols (pre-terminals). Obviously, these two problems are closely

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 187 of 404

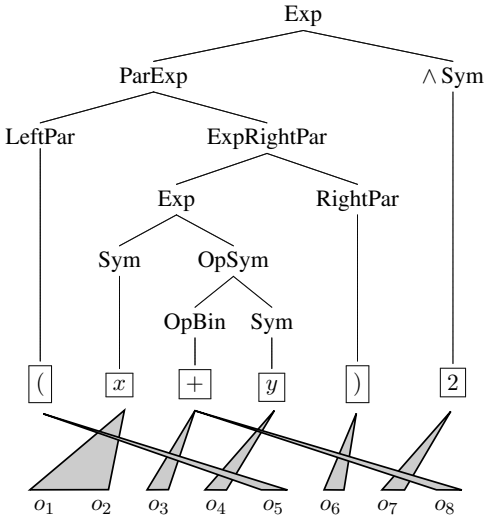


Figure 12. Parse tree of expression $(x + y)^2$ given the input sequence of strokes described in Figure 11. The parse tree represents the structure of the mathematical expression and it produces the 6 recognized symbols that account for the 8 input strokes.

related. Symbol recognition is influenced by the structure of the mathematical expression, and detecting the structure of the math expression strongly depends on the segmentation and recognition of symbols. For these reasons we propose an integrated strategy that computes the most likely parse tree while simultaneously solving symbol segmentation, symbol recognition and the structural analysis of the input.

3.1. Statistical Framework

Formally, let a mathematical expression be a sequence of N strokes $\boldsymbol{o} = o_1 o_2 \dots o_N$. We pose the mathematical expression recognition as a structural parsing problem where the goal is to obtain the most likely parse tree t that accounts for the input sequence of strokes $\boldsymbol{o}$:

$$\hat{t} = \arg\max_{t \in T} p(t \mid \mathbf{o})$$

where $\mathcal{T}$ represents the set of all possible parse trees.

At this point, we consider the sequence of mathematical symbols $\boldsymbol{s} \in \mathcal{S}$ as a hidden variable, where $\mathcal{S}$ is the set of all possible sequences of symbols (pre-terminals) produced by the parse tree t: $\boldsymbol{s} = \text{yield}(t)$. This can be formalized as follows:

$$\hat{t} = \arg\max_{t \in T} \sum_{\substack{\boldsymbol{s} \in \mathcal{S} \\ \boldsymbol{s} = \text{yield}(t)}} p(t, \boldsymbol{s} \mid \boldsymbol{o})$$

If we approximate the previous probability by the maximum probability parse tree, and assume that the structural part of the equation depends only on the sequence of pre-terminals

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 188 of 404

$\boldsymbol{s}$, the target expression becomes

$$\hat{t} \approx \arg\max_{t \in T} \max_{\substack{\boldsymbol{s} \in \mathcal{S} \\ \boldsymbol{s}=\text{yield}(t)}} p(\boldsymbol{s} \mid \boldsymbol{o}) \cdot p(t \mid \boldsymbol{s}) \quad (1)$$

such that $p(\boldsymbol{s}|\boldsymbol{o})$ represents the observation (symbol) likelihood and $p(t|s)$ represents the structural probability.

This problem can be solved in two steps. First, by calculating the segmentation of the input into mathematical symbols and, second, by computing the structure that relates all recognized symbols (Zanibbi et al., 2002).

However, we propose here a fully integrated strategy for computing Equation (1) where symbol segmentation, symbol recognition and structural analysis of the input expression are globally determined. This way, all the information is taken into account in order to obtain the most likely mathematical expression.

In Section below we define the observation model that accounts for the probability of recognition and segmentation of symbols, $p(s|o)$. The probability that accounts for the structure of the mathematical expression $p(t|s)$ is described in the Structural Probability Section.

3.2. Symbol Likelihood

As we have seen in the recognition of on-line handwritten math expressions, the input is a sequence of strokes $o = o_1 o_2 \ldots o_N$, which encodes a sequence of pre-terminals $s = s_1 s_2 \ldots s_K$, $(1 \leq K \leq N)$ that represents the mathematical symbols. A symbol is made up of one or more strokes. Some approaches used have assumed that users always write a symbol with consecutive strokes (Shi et al., 2007; Yamamoto et al., 2006). Although this assumption may be true in many cases, it constitutes a severe constraint that means that these models cannot account for symbols composed of non-consecutive written strokes. For example, the plus sign $(+)$ in the expression in Figure 11 is made up of strokes o_3 and o_8 and would not be recognized by a model that incorporates this assumption.

In this section we define a symbol likelihood model that is not based on time information but rather spatial information. This model is therefore able to recognize mathematical symbols made up of non-consecutive strokes. Given a sequence of strokes, testing all possible segmentations could be unfeasible given the high number of possible combinations. However, it is clear that only strokes that are close together will form a mathematical symbol, which is why we tackle the problem using the spatial and geometric information available since, by doing so, we can effectively reduce the number of symbol segmentations considered. The application of this intuitive idea is detailed in the next section.

Before defining the segmentation strategy adopted for modeling the symbol likelihood, we must introduce some preliminary formal definitions.

Definition 3. *Given a sequence of N input strokes o, and the set containing them* $\text{set}(o) = \{o_i \mid i : 1 \ldots N\}$, *a* segmentation *of o into K segments is a partition of the set of input strokes*

$$b(o, K) = \{\, b_i \mid i : 1 \ldots K \,\}$$

where each b_i is a set of (possibly non-consecutive) strokes representing a segmentation hypothesis for a given symbol.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 189 of 404

Definition 4. *We define $\mathcal{B}_K$ as the set of all possible segmentations of the input strokes o in K parts. Similarly, we define the set of all segmentations $\mathcal{B}$ as:*

$$\mathcal{B} = \bigcup_{1 \leq K \leq N} \mathcal{B}_K$$

Then, in Equation (1), we can define a generative model $p(s, o)$, rather than $p(s|o)$, because, given that the term $p(o)$ does not depend on the maximization variables s and t, we can drop it. The next step is to replace the sequence of N input strokes o by its previously defined set of segmentations, $b = b(o, K) \in \mathcal{B}_K$ where $1 \leq K \leq N$. Finally, given K, we define a hidden variable that limits the number of strokes for each of the K pre-terminals (symbols) that make up the segmentation, $l : l_1 \ldots l_K$. Each l_i falls within the range $1 \leq l_i \leq \min(N, L_{\max})$, where $L_{\max}$ is a parameter that constrains the maximum number of strokes that a symbol can have.

$$p(s, o) = \sum_{1 \leq K \leq N} \sum_{b \in \mathcal{B}_K} \sum_{l} p(s, b, l)$$

In order to develop this expression, we factor it with respect to the number of pre-terminals (symbols) and assume the following constraints: 1) we approximate the summations by maximizations, 2) the probability of a possible segmentation depends only on the spatial constraints of the strokes it is made up of, 3) the probability of a symbol depends only on the set of strokes associated with it, and 4) the number of strokes for a pre-terminal depends only on the symbol it represents:

$$p(s, o) \approx \max_{K} \max_{b \in \mathcal{B}_K} \max_{l} \prod_{i=1}^{K} p(b_i)\, p(s_i \mid b_i)\, p(l_i \mid s_i) \tag{2}$$

From Equation (2) we can conclude that we need to define three models: a *symbol segmentation model*, $p(b_i)$, a *symbol classification model*, $p(s_i|b_i)$, and a *symbol duration model*, $p(l_i|s_i)$.

3.2.1. Symbol Segmentation Model

Many symbols in mathematical expressions are made up of more than one stroke. For example, the symbols x and $+$ in Figure 11 have two strokes, while symbols like π or $\neq$ usually require three strokes, etc. As we have already discussed, in this section we are proposing a model where stroke segmentation is not based on temporal information, but rather on spatial and geometric information. We also defined $\mathcal{B}$ as the set of all possible segmentations. Given this definition of $\mathcal{B}$, it is easy to see that its size is exponential on the number of strokes N. In this section we first explain how to effectively reduce the number of segmentations considered, and then we describe the segmentation model used for computing the probability of a certain hypothesis $p(b_i)$.

Given a mathematical expression represented by a sequence of strokes o, the number of all possible segmentations $\mathcal{B}$ could be unfeasible. In order to reduce this set, we use two concepts based on geometric and spatial information: *visibility* and *closeness*. Let us first introduce some definitions.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 190 of 404

Definition 5. *The distance between two strokes o_i and o_j can be defined as the Euclidean distance between their closest points.*

Definition 6. *A stroke o_i is considered* visible *from o_j if the straight line between the closest points of both strokes does not cross any other stroke o_k.*

If a stroke o_i is not *visible* from o_j we consider that their distance is infinite. For example, given the expression in Figure 11, the strokes visible from o_4 are o_3, o_6 and o_8.

Furthermore, we know that a multi-stroke symbol is composed of strokes that are spatially close. For this reason, we only consider segmentation hypotheses b_i where strokes are close to each other.

Definition 7. *A stroke o_i is considered* close *to another stroke o_j if their distance is shorter than a given threshold.*

Using these definitions, we can characterize the set of possible segmentation hypotheses.

Definition 8. *Let G be an undirected graph such that each stroke is a node and edges only connect strokes that are* visible *and* close*. Then, a segmentation hypothesis b_i is admissible if the strokes it contains form a connected subgraph in G.*

Consequently, a segmentation $b(o, K) = b_1 b_2 \dots b_K$ is admissible if each b_i is, in turn, admissible. These two geometric and spatial restrictions significantly reduce the number of possible symbol segmentations.

We need a segmentation model in order to calculate the probability that a given set of strokes (segmentation hypothesis, b_i) forms a mathematical symbol. Commonly, symbol segmentation models are defined using different features based on geometric information (Lehmberg et al., 1996). Also, the shape of the hypotheses has been used (Hu and Zanibbi, 2013).

In this proposal, we used a segmentation model very similar to the concept of *grouping likelihood* proposed in Shi et al. (2007). As in Shi et al. (2007), we defined a set of geometric features associated with a segmentation hypothesis b_i. First, for each stroke o_j of b_i, we calculated the mean horizontal position, the mean vertical position and its size computed as the maximum value of horizontal and vertical size. Then, for each pair of strokes we calculated the difference between their horizontal positions, vertical positions and sizes. The average of these differences for each pair determined the features used for the segmentation model: *average horizontal distance* (d), *average vertical offset* (σ), and *average size difference* (δ). Additionally, we defined another feature: *average distance* (θ). This last feature is computed as the distance between the closest points of two strokes.

The authors in (Shi et al., 2007) used a scoring function where that these features were normalized using a fixed threshold value. However, this normalization depends on the resolution of the input. In order to overcome this restriction we normalized the features by the diagonal of the *normalized symbol size* (see the Complexity and Search Space Section), thereby ensuring that features are resolution-independent.

Finally, instead of the scoring function proposed in Shi et al. (2007), we trained a Gaussian Mixture Model (GMM) using positive samples $c = 1$ (the strokes of b_i can form a mathematical symbol) and a GMM using negative samples $c = 0$ (the strokes of b_i

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 191 of 404

cannot form a mathematical symbol) from the set of all admissible segmentations $\mathcal{B}$. A segmentation hypothesis b_i is represented by the 4-dimensional normalized feature vector $\mathrm{g}(b_i) = [d, \sigma, \delta, \theta]$, and the probability $p(b_i)$ that a hypothesis b_i forms a mathematical symbol is obtained as

$$p(b_i) = p_{\mathrm{GMM}}(c = 1 \mid \mathrm{g}(b_i)) \tag{3}$$

3.2.2. Symbol Classification Model

Symbol classification is crucial in order to properly recognize mathematical notation. In this section we describe the model used for calculating the probability that a certain segmentation hypothesis b_i represents a mathematical symbol s_i, i.e. the probability $p(s_i|b_i)$ required in Equation (2).

Several approaches have been proposed in the literature to tackle this problem using different classifiers: Artificial Neural Networks (Thammano and Rugkunchon, 2006), Support Vector Machines (SVM) (Keshari and Watt, 2007), Gaussian Mixture Models (GMM) (Shi et al., 2007), elastic matching (MacLean and Labahn, 2010), Hidden Markov Models (HMMs) (Winkler, 1996; Hu and Zanibbi, 2011) and Recurrent Neural Networks (RNN) (Álvaro et al., 2013). Although not all of these approaches have been tested since some publications used private datasets, Bidirectional Long Short-Term Memory RNNs (BLSTM-RNN) are a state-of-the-art model that has outperformed previously reported results (Álvaro et al., 2013). For this reason we used a BLSTM-RNN for mathematical symbol classification.

RNNs are a connectionist model containing a self-connected hidden layer. The recurrent connection provides information about previous inputs, meaning that the network can benefit from past contextual information (Pearlmutter, 1989). Long Short-Term Memory (LSTM) is an advanced RNN architecture that allows cells to access context information over long periods of time. This is achieved by using a hidden layer made up of recurrently connected subnets called memory blocks (Graves et al., 2009).

Bidirectional RNNs (Schuster and Paliwal, 1997) have two separate hidden layers that allow the network to access context information in both time directions: one hidden layer processes the input sequence forwards while another processes it backwards. The combination of bidirectional RNNs and the LSTM architecture results in BLSTM-RNNs, which have outperformed standard RNNs and HMMs in handwriting text recognition (Graves et al., 2009) and handwritten mathematical symbol classification (Álvaro et al., 2013). They are also faster than HMMs in terms of classification speed.

In order to train a BLSTM-RNN classifier, we computed several features from a segmentation hypothesis. Given a mathematical symbol represented as a sequence of points, for each point $p = (x, y)$ we extracted the following 7 on-line features:

- Normalized coordinates: (x, y) normalized values such that $y \in [0, 100]$ and the aspect-ratio of the sample is preserved.
- Normalized first derivatives: (x', y').
- Normalized second derivatives: (x'', y'').
- Curvature: k, the inverse of the radius of the curve at each point.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 192 of 404

It should be noted that no resampling is required prior to the feature extraction process because first derivatives implicitly perform writing speed normalization (Toselli et al., 2007).

Furthermore, the combination of on-line and off-line information has been proven to improve recognition accuracy (Winkler, 1996; Keshari and Watt, 2007; Álvaro et al., 2014). For this reason, we also rendered the image representing the symbol hypothesis b_i and extracted off-line features to train another BLSTM-RNN classifier.

Following (Álvaro et al., 2014, 2016), for a segmentation hypothesis b_i, we generated the image representation as follows. We set the image height to H pixels and kept the aspect ratio (up to $5H$, in order to prevent creating images that were too wide). Then we rendered the image representation by using linear interpolation between each two consecutive points in a stroke. The final image was produced after smoothing it using a mean filter with a window sized 3×3 pixels, and binarizing for every pixel that is different from the background (white).

Given a binary image of height H and W columns, for each column we computed 9 off-line features (Marti and Bunke, 2001; Álvaro et al., 2014):

- Number of black pixels in the column.
- Center of gravity of the column.
- Second order moment of the column.
- Position of the upper contour of the column.
- Position of the lower contour of the column.
- Orientation of the upper contour of the column.
- Orientation of the lower contour of the column.
- Number of black-white transitions in the column.
- Number of black pixels between the upper and lower contours.

In order to classify a mathematical symbol hypothesis, we trained two classifiers: a BLSTM-RNN with on-line feature vectors, and a BLSTM-RNN with off-line feature vectors. The BLSTM-RNN was trained using a frame-based approach. Given a symbol hypothesis b_i of n frames, we computed a sequence of n feature vectors. Then, we obtained the posterior probability per symbol normalized as its average probability per frame:

$$p(s \mid b_i) = \frac{1}{n} \sum_{j=1}^{n} p(s \mid f_j) \tag{4}$$

Finally, given a segmentation hypothesis b_i and using Equation (4), we obtained the posterior probability of a BLSTM-RNN with on-line features and the posterior probability of a BLSTM-RNN with off-line features. We combined the probabilities of both classifiers using linear interpolation and a weight parameter (α). The final probability of the symbol classification model is calculated as

$$p(s_i \mid b_i) = \alpha \cdot p_{\text{on}}(s_i \mid b_i) + (1 - \alpha) \cdot p_{\text{off}}(s_i \mid b_i) \tag{5}$$

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 193 of 404

3.2.3. Symbol Duration Model

The symbol duration model accounts for the intuitive idea that a mathematical symbol class is usually made up of a certain number of strokes. For example, the plus sign (+) is likely to be composed of two strokes, rather than one or more than two strokes.

As authors proposed in (Shi et al., 2007), a simple way to calculate the probability that a certain symbol class s_i is made up of l_i strokes is

$$p(l_i \mid s_i) = \frac{c(s_i, l_i)}{c(s_i)} \tag{6}$$

where $c(s_i, l_i)$ is the number of times the symbol s_i was composed of l_i strokes and $c(s_i)$ is the total number of samples of class s_i in the set used for estimation. We smoothed these probabilities in order to account for unseen events.

3.3. Structural Probability

The proposed statistical framework raises the problem of recognizing a mathematical expression as finding the most likely parse tree t that accounts for the input strokes o. Formally, the problem is stated in Equation (1) such that two probabilities are required. In the previous section we presented the calculation of the *symbol likelihood* $p(s|o)$. In this section we will define the *structural probability* $p(t|s)$.

Although the most natural way to compute the most likely parse tree of an input sequence would be to define probabilistic *parsing models* $p(t|s)$, in the literature, this problem has usually been tackled using generative models $p(t, s)$ (*language models*) and, more precisely, *grammatical models* (Manning and Schütze, 1999). Next we define a generative model $p(t, s)$ based on a two-dimensional extension of the well-known context-free grammatical models.

3.3.1. 2D Probabilistic Context-Free Grammars

A context-free model is a powerful formalism able to represent the structure of natural languages. It is an appropriate model to account for mathematical notation given the structural dependencies existing between the different elements in an expression (for instance, the parentheses in Figure 11). We will use a two-dimensional extension of PCFG, a well-known formalism widely used for mathematical expression recognition (Anderson, 1967; Chou, 1989; Yamamoto et al., 2006; Awal et al., 2014; Álvaro et al., 2014).

Definition 9. *A* Context-Free Grammar *(CFG) G is a four-tuple $(\mathcal{N}, \Sigma, S, \mathcal{P})$, where $\mathcal{N}$ is a finite set of non-terminal symbols, Σ is a finite set of terminal symbols ($\mathcal{N} \cap \Sigma = \emptyset$), $S \in \mathcal{N}$ is the start symbol of the grammar, and $\mathcal{P}$ is a finite set of rules: $A \to \alpha$, $A \in N$, $\alpha \in (\mathcal{N} \cup \Sigma)^+$.*

A CFG in Chomsky Normal Form (CNF) is a CFG in which the rules are of the form $A \to BC$ or $A \to a$ (where $A, B, C \in \mathcal{N}$ and $a \in \Sigma$).

Definition 10. *A* Probabilistic CFG *(PCFG) $\mathcal{G}$ is defined as a pair (G, p), where G is a CFG and $p : P \to]0, 1]$ is a probability function of rule application such that $\forall A \in \mathcal{N} : \sum_{i=1}^{n_A} p(A \to \alpha_i) = 1$, where n_A is the number of rules associated with non-terminal symbol A.*

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 194 of 404

Definition 11. *A* Two-Dimensional PCFG *(2D-PCFG) is a generalization of a PCFG, where terminal and non-terminal symbols describe two-dimensional regions. This grammar in CNF results in two types of rules: terminal rules and binary rules. First, the terminal rules $A \rightarrow a$ represent the mathematical symbols which are ultimately the terminal symbols of 2D-PCFG. Second, the binary rules $A \xrightarrow{r} BC$ have an additional parameter r that represents a given spatial relationship, and its interpretation is that regions B and C must be spatially arranged according to the spatial relationship r.*

In the Spatial Relationships Model Section we will provide a full description of the spatial relationships considered here in order to address the recognition of mathematical expressions. The construction of the 2D-PCFG and the estimation of the probabilities are detailed in the 2D-PCFG Estimation Section.

3.3.2. Parse Tree Probability

The 2D-PCFG model allows us to calculate the structural probability of a mathematical expression in terms of the joint probability $p(t, s)$, so that in CNF it is computed as:

$$p(t,s) = \prod_{(A \rightarrow a, t)} p(a \mid A) \prod_{(A \rightarrow BC, t)} p(BC \mid A)$$

where $p(\alpha|A)$ is the probability of the rule $A \rightarrow \alpha$ and represents the probability that α is derived from A. Moreover, $(A \rightarrow \alpha, t)$ denotes all rules $(A \rightarrow \alpha)$ contained in the parse tree t. In the defined 2D extension of PCFG, the composition of subproblems has an additional constraint according to a spatial relationship r. Let the spatial relationship r between two regions be a hidden variable. Then, the probability of a binary rule is written as:

$$p(BC \mid A) = \sum_{r} p(BC, r \mid A)$$

When the inner probability in the previous addition is estimated from samples, the mode is the dominant term. Therefore, by approximating summations by maximizations, and assuming that the probability of a spatial relationship depends only on the subproblems B and C involved, the structural probability of a mathematical expression becomes:

$$p(t,s) \approx \prod_{(A \rightarrow a, t)} p(a \mid A) \tag{7}$$

$$\prod_{(A \rightarrow BC, t)} \max_{r} p(BC \mid A)\, p(r \mid BC) \tag{8}$$

where $p(a|A)$ and $p(BC|A)$ are the probabilities of the rules of the grammar, and $p(r|BC)$ is the probability that regions encoded by non-terminals B and C are arranged according to spatial relationship r.

3.3.3. Spatial Relationships Model

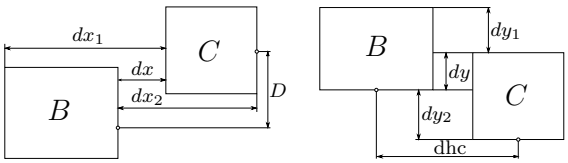
The definition of Equation (8) for computing the structural probability of a mathematical expression requires a spatial relationship model. This model provides the probability $p(r|BC)$ that two subproblems B and C are arranged according to spatial relationship r.

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 195 of 404

A common approach for obtaining a spatial relationship model is to define a set of geometric features to train a statistical classifier. Most proposals in the literature define geometric features based on the bounding boxes of the regions (Zanibbi et al., 2002; Álvaro et al., 2014; Awal et al., 2014), although a proposal based on shape descriptors has also been studied (Álvaro and Zanibbi, 2013). The geometric features are usually modeled using Gaussian models (Awal et al., 2014), SVM (Álvaro et al., 2014) or fuzzy functions (Zhang et al., 2005), though some authors manually define specific functions (Zanibbi et al., 2002; Yamamoto et al., 2006; MacLean and Labahn, 2013).

In this work, we deal with the recognition of mathematical expressions using six spatial relationships: *right* (BC), *below* ($\frac{B}{C}$), *subscript* (B_C), *superscript* (B^C), *inside* ($\sqrt{C}$) and *mroot* ($\sqrt[C]{\ }$).

In order to train a statistical classifier, given two regions B and C we define nine geometric features based on their bounding boxes (Álvaro and Zanibbi, 2013) (see Figure 13). This way, we compute the feature vector $\mathrm{h}(B, C)$ that represents their relationship and can be used for classification. The features are defined in Figure 13, where H is the height of region C, feature D is the difference between the vertical centroids, and dhc is the difference between the horizontal centers. The features are normalized by the combined height of regions B and C. The most challenging classification is between classes *right*, *subscript*



$$\mathrm{h}(B, C) = [H, D, \mathrm{dhc}, dx, dx_1, dx_2, dy, dy_1, dy_2]$$

Figure 13. Geometric features for classifying the spatial relationship between regions B and C.

and *superscript* (Álvaro et al., 2014; Álvaro and Zanibbi, 2013). An important feature for distinguishing between these three relationships is the difference between vertical centroids (D). Some symbols have ascenders, descenders or certain shapes where that the vertical centroid is not the best placement for the symbol center.

With a view to improving the placement of vertical centroids, we divided symbols into four typographic categories: ascendant (e.g. d or λ), descendant (p, μ), normal ($x, +$) and middle ($7, \Pi$). For *normal* symbols the centroid is set to the vertical centroid. For *ascendant* symbols the centroid is shifted downward to $(centroid + bottom)/2$. Likewise, for *descendant* symbols the centroid is shifted upward to $(centroid + top)/2$. Finally, for *middle* symbols the vertical centroid is defined as $(top + bottom)/2$.

Once we defined the feature vector representing a spatial relationship, we can train a GMM using labeled samples so that the probability of the spatial relationship model can be computed as the posterior probability provided by the GMM for class r

$$p(r \mid BC) = p_{\mathrm{GMM}}(r \mid \mathrm{h}(B, C))$$

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 196 of 404

This model is able to provide a probability for every spatial relationship r between any two given regions. However, there are several situations where we would not want the statistical model to assign the same probability as in other cases. Considering the expression in Figure 14, the GMMs might yield a high probability for *superscript* relationship '3^x', for the *below* relationship '$\frac{\pi}{2}$', and for the *right* relationship '2 3'; though we might expect a lower probability, since they are not the true relationships in the correct mathematical expression.

Intuitively, those symbols or subexpressions that are closer together should be combined first. Furthermore, two symbols or subexpressions that are not *visible* from each other should not be combined. These ideas are introduced into the spatial relationship model as a penalty based on the distance between strokes.

Specifically, given the combination of two hypotheses B and C, we computed a penalty function based on the minimum distance between the strokes of B and C

$$\text{penalty}(B, C) = 1/(\ 1 + \min_{o_i \in B,\, o_j \in C} \text{d}(o_i, o_j)\ )$$

so that it is in the range $[0, 1]$. It should be noted that, although it is a penalty function, since it multiplies the probability of a hypothesis, the lower the penalty value is, the greater the probability is penalized.

This function is based on the single-linkage hierarchical clustering algorithm (Sibson, 1973) where, at each step, the two clusters separated by the shortest distance are combined. We defined a penalty function in order to avoid making hard decisions, because it is not always the case that the two closest strokes must be combined first.

The final statistical spatial relationship probability is computed as the product of the probability provided by the GMM and the penalty function based on hierarchical clustering

$$p(r \mid BC) = p_{\text{GMM}}(r \mid \text{h}(B, C)) \cdot \text{penalty}(B, C) \tag{9}$$

An interesting property of the application of the penalty function is that, given that the distance between non-visible strokes is considered infinite, this function prunes many hypotheses. Furthermore, it favors the combination of closer strokes over strokes that are further apart. For example, in the superscript relationship between symbols 3 and x in Figure 14, although it could be likely, the penalty will favor that the 3 is first combined with the fraction bar, and later the fraction bar (and the entire fraction) with the x.

3.4. Parsing Algorithm

In this section we present the parsing algorithm for mathematical expression recognition that maximizes Equation (1). We define a CYK-based algorithm for 2D-PCFGs in the

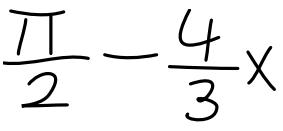


Figure 14. Example for hierarchical clustering penalty.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 197 of 404

statistical framework described previously. Using this algorithm, we compute the most likely parse tree according to the proposed model.

The parsing algorithm is essentially a dynamic programming method. First, the initialization step computes the probability of several mathematical symbols for each possible segmentation hypothesis. Second, the general case computes the probability of combining different hypotheses so that it builds the structure of the mathematical expression.

The dynamic programming algorithm computes a probabilistic parse table γ. Following a notation similar to (Goodman, 1999), each element of γ is a probabilistic non-terminal vector, where their components are defined as:

$$\gamma(A, b, l) \;=\; \hat{p}(A \stackrel{*}{\Rightarrow} b); \qquad l \;=\mid b \mid$$

where that $\gamma(A, b, l)$ denotes the probability of the best derivation that the non-terminal A generates a set of strokes b of size l.

Initialization: In this step the parsing algorithm computes the probability of every admissible segmentation $b \in \mathcal{B}$ as described in the Symbol Segmentation Model Section. The probability of each segmentation hypothesis is computed according to Eqs. (1) and (2) as

$$\gamma(A, b_i, l) = \max_{s} \{\, p(s \mid A)\; p(b_i)\; p(s \mid b_i)\; p(l \mid s) \,\} \qquad (10)$$

$$\forall A, \forall K,\ \forall b \in \mathcal{B}_K,\ 1 \leq i \leq |b|,\ 1 \leq l \leq \min(N, L_{\max})$$

where $L_{\max}$ is a parameter that constrains the maximum number of strokes that a symbol can have.

This probability is the product of a range of factors so that it is maximized for every mathematical symbol class s: probability of terminal rule, $p(s|A)$ (Equation (7)), probability of segmentation model, $p(b)$ (Equation (3)), probability of mathematical symbol classifier, $p(s|b)$ (Equation (5)), and probability of duration model probability, $p(l|s)$ (Equation (6)).

General case: In this step the parsing algorithm computes a new hypothesis $\gamma(A, b, l)$ by merging previously computed hypotheses from the parsing table until all N strokes are parsed. The probability of each new hypothesis is calculated according to Eqs. (1) and (8) as:

$$\gamma(A, b, l) = \max\{\, \gamma(A, b, l),\ \max_{B,C}\ \max_{r}\ \max_{b_B, b_C} \{ \\ p(BC \mid A)\gamma(B, b_B, l_B)\, \gamma(C, b_C, l_C)\, p(r \mid BC)\ \}\} \qquad (11)$$

$$\forall A,\ 2 \leq l \leq N$$

where $b = b_B \cup b_C;\ b_B \cap b_C = \emptyset$ and $l = l_B + l_C$.

This expression shows how a new hypothesis $\gamma(A, b, l)$ is built by combining two subproblems $\gamma(B, b_B, l_B)$ and $\gamma(C, b_C, l_C)$, considering both syntactic and spatial information: probability of binary grammar rule $p(BC|A)$ (Equation (8)) and probability of spatial relationship classifier $p(r|BC)$ (Equation (9)). It should be noted that both distributions significantly reduce the number of hypotheses that are merged. Also, the probability is

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 198 of 404

maximized taking into account that a probability might already have been set by the Equation (10) during the initialization step.

Finally, the most likely hypothesis and its associated derivation tree $\hat{t}$ that accounts for the input expression can be retrieved in $\gamma(S, o, N)$ (where S is the start symbol of the grammar).

3.4.1. Complexity and Search Space

We have defined an integrated approach for math expression recognition based on parsing 2D-PCFG. The dynamic programming algorithm is defined by the corresponding recursive equations. The initialization step is performed by Equation (10), while the general case is computed according to Equation (11). In addition to the formal definition, there are some details of the parsing algorithm regarding the search space that need further explanation.

Once several symbol hypotheses have been created during the initialization step, the general case is the core of the algorithm where hypotheses of increasing size $2 \leq l \leq N$ are generated with Equation (11). For a given size l, we have to test all the sizes in order to split l into hypotheses b_B and b_C so that $l = l_B + l_C$. Once the sizes are set, for every set of strokes b_B we have to test every possible combination with another set b_C using the binary rules of the grammar $A \xrightarrow{r} BC$. According to this, we can see that the time complexity for parsing an input expression of N strokes is $O(N^4|P|)$ where $|P|$ is the number of productions of the grammar. However, this complexity can be reduced by constraining the search space.

The intuitive idea is that, given a set of strokes b_B, we do not need to try to combine it with every other set b_C. A set of strokes b_B defines a region in space, allowing us to limit the set of hypothesis b_C to those that fall within a region of interest. For example, given symbol 4 in Figure 14, we only have to check for combinations with the fraction bar and symbol 3 (below relationship) and the symbol x (right or sub/superscript relationships).

We applied this idea as follows. Given a stroke o_i we define its associated region $r(o_i) = (x, y, s, t)$ in the 2D space as the minimum bounding box that contains that stroke, where (x, y) is the top-left coordinate and (s, t) the bottom-right coordinate of the region. Likewise, given a set of strokes $b = \{o_j \mid 1 \leq j \leq |b|\}$ we define $r(b) = (x_b, y_b, s_b, t_b)$ as the minimum rectangle that contains all the strokes $o_j \in b$. Therefore, given a spatial region $r(b_B)$ we retrieve only the hypotheses b_C whose region $r(b_C)$ falls in a given area $\mathcal{R}$ relative to $r(b_B)$. Figure 15 shows the definition of the regions in the space in order to retrieve relevant hypotheses to combine with b_B depending on the spatial relation. The dimensions of the normalized symbol size (R_w, R_h) are computed as: R_w, the maximum between the average and median width of the input strokes; and R_h, the maximum between the average and median height of the input strokes. These calculations are independent of the input resolution. The normalized symbol size is also used to normalize other distance-related metrics in the model, like determining what strokes are close together in the multi-stroke symbol recognition or the normalization factor of features in the segmentation model.

In order to efficiently retrieve the hypotheses falling in a given region $\mathcal{R}$, every time a set of hypotheses of size l_A is computed, we sort this set according to the x coordinate of every region $r(b_A)$ associated with $\gamma(A, b_A, l_A)$. This sorting operation has cost $O(N \log N)$. Afterwards, given a rectangle $r(b_B)$ in the search space and a size l_C, we can retrieve the

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 199 of 404

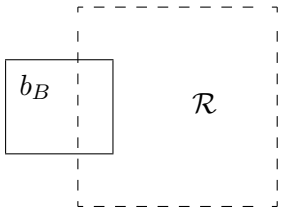
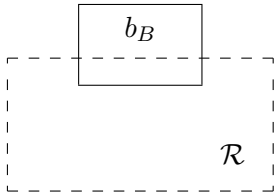
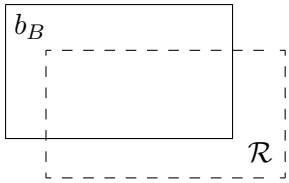
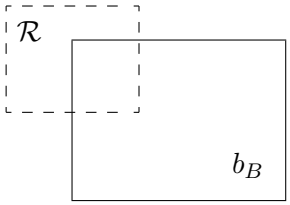
Right, Sub/Superscript $x = \max(r(b_B).x + 1,$ $\quad r(b_B).s - R_w)$ $y = r(b_B).y - R_h$ $s = r(b_B).s + 8R_w$ $t = r(b_B).t + R_h$	
Below $x = r(b_B).x - 2R_w$ $y = \max(r(b_B).y + 1,$ $\quad r(b_B).t - R_h)$ $s = r(b_B).s + 2R_w$ $t = r(b_B).t + 3R_h$	
Inside $x = r(b_B).x + 1$ $y = r(b_B).y + 1$ $s = r(b_B).s + R_w$ $t = r(b_B).t + R_h$	
Mroot $x = r(b_B).x - 2R_w$ $y = r(b_B).y - R_h$ $s = \min(r(b_B).s,$ $\quad r(b_B).x + 2R_w)$ $t = \min(r(b_B).t,$ $\quad r(b_B).y + 2R_h)$	

Figure 15. Spatial regions defined to retrieve hypotheses relative to hypothesis b_B according to different relations.

hypotheses $\gamma(C, b_C, l_C)$ falling within that area by performing a binary search over that set in $O(\log N)$. Although the regions are arranged in two-dimensions and they are sorted only in one dimension, this approach is reasonable since mathematical expressions grow mainly from left to right.

Assuming that this binary search will retrieve a small constant number of hypothesis, the final complexity achieved is $O(N^3 \log N|P|)$. Furthermore, many unlikely hypotheses are pruned during the parsing process.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 200 of 404

4. The Problem of Performance Evaluation

Assessing the performance of different solutions to a problem is crucial in order to evaluate the advancements in a specific field so that research can move towards the best approach to deal with it. A good set of performance metrics along with large public datasets is the desired scenario for comparing different approaches and helping the research community. Unbiased metrics that can be computed automatically are very important for objective evaluation. Furthermore, in many pattern recognition problems it is common to estimate the parameters of a model by minimizing an error function based on a certain metric.

Automatic performance evaluation in mathematical expression recognition is not straightforward. There are several issues that have made comparison in this field difficult.

A deep discussion about this problem can be found in Lapointe and Blostein (2009) and Awal et al. (2010). Next, we review the main problems that make automatic performance evaluation difficult in this field.

One of the main issues in performance evaluation of math notation is that there are many ambiguities at different levels. First, there are ambiguities inherent to the expressions that accept different interpretations. Awal et al. (2010) show some examples like the expression $f(y+1)$ that can be considered as the variable f multiplying the term $(y+1)$, or the function f applied to the value $y+1$; or the expression $a/2b$ that can be interpreted as a fraction with denominator $2b$ or the product between the fraction $a/2$ and the variable b. Other ambiguities are due to handwriting production. In the first sections there are several examples of ambiguities at different levels, for example Figure 4 shows ambiguous symbol segmentations and Figure 5 presents different interpretations of the same shapes.

These previous sources of ambiguity demonstrate that more than one ground-truth could be valid for a given expression. Nevertheless, even if a math expression is not ambiguous, the representation formats do not enforce uniqueness (Lapointe and Blostein, 2009). Math expressions are usually encoded in LaTeX or MathML, where the same expression can be annotated by several correct representations as shown in Figure 16. All the described ambiguities can result in a correct recognition result for a given expression not matching the ground-truth, thereby reporting undesired recognition errors. Consequently, metrics for automatic performance evaluation of math expression recognition should be based on formats that specify a unique encoding for a given math expression.

Commonly, mathematical expression recognition is divided into three different problems (see the first section): segmentation, symbol recognition and structural analysis. Although several issues have been described, symbol segmentation and symbol recognition can be easily calculated. The only remaining ambiguity is the interpretation of an expression. Measuring errors in the structure of the expression is the most challenging task. Many authors report symbol segmentation rate, symbol recognition rate and the expression recognition rate. However, the expression recognition rate is hard to automate due to the representation ambiguities, thus several results are computed manually (Awal et al., 2010). Global error values can be computed as an edit distance between strings or trees, but the encoding of the math expressions has to deal with the representation ambiguities. Furthermore, edit distances report a global error (frequently not normalized), where the source of the error is unknown (segmentation, symbols, structure). In the following sections we detail several proposals of metrics for automatic evaluation of math expression recognition,

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 201 of 404

LaTeX	MathML	
x_a^2 + 1 x_a^{2} + 1 x_{a}^2 + 1 x_{a}^{2} + 1 x^2_a + 1 x^2_{a} + 1 x^{2}_a + 1 x^{2}_{a} + 1	<mathml> <mrow> <msubsup> <mi>x</mi> <mi>a</mi> <mn>2</mn> </msubsup> <mo>+</mo> <mn>1</mn> </mrow> </mathml>	<mathml> <mrow> <msubsup> <mi>x</mi> <mi>a</mi> <mn>2</mn> </msubsup> <mrow> <mo>+</mo> <mn>1</mn> </mrow> </mrow> </mathml>

Figure 16. Some examples of different valid representations for math expression $x_a^2 + 1$ in LaTeX and MathML format.

analyzing their strengths and weaknesses.

4.1. Early Global Metrics

Expression recognition rate is a metric for computing the overall performance of a math expression recognition system. It is commonly reported along with other metrics at symbol level (Okamoto et al., 2001; Zanibbi et al., 2002). Recognition rate at expression level complements the symbol level evaluation, but it is a pessimistic metric because a single error causes the entire expression to be a wrong recognition result. Furthermore, its computation has to deal with representation ambiguities.

Later, other global metrics were proposed as a combination of recognition rates at different levels. Chan and Yeung (2001) proposed an *integrated performance measure* as the ratio of the number of correctly recognized symbols and operators (structure) to the total number of symbols and operators tested. Garain and Chaudhuri (2005) defined a *global performance index* that combines the number of symbols recognized incorrectly and the number of symbols incorrectly arranged in the expression. They also penalized differently the structural errors depending on the *level* of the symbol, so that the dominant baseline of an expression is treated as level zero and the level number increases above and decreases below the baseline.

These first proposals for computing a global error integrate the errors at symbol level and at structural level. However, segmentation errors are not taken into account and would affect the computation of these metrics because indirect matching could be possible between expressions. Also, determining implicit operators in the *integrated performance measure* or the incorrect arrangements in levels of the *global performance index* is not straightforward, and the software for evaluation was not made available.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 202 of 404

4.2. EMERS

A mathematical expression can be naturally represented as a tree (see Figure 7). The tree representation, commonly in MathML format, contains simultaneously the symbols and the structure of a given mathematical expression. For this reason, computing an edit distance between trees is an appropriate method in order to compute the error between a recognized expression and its ground-truth tree.

Sain et al. (2010) proposed EMERS,[1] a tree matching-based performance evaluation metric for mathematical expression recognition. Using the tree representation of two expressions in MathML (which can also be easily obtained from LaTeX) they defined a method for computing the edit distance between them. Since matching of trees is a hard problem, they proposed to match ordered trees represented by their corresponding Euler strings. Given two trees encoded by two Euler strings A and B, the overall complexity of the EMERS algorithm is $O(|A|^2|B|^2)$ or more generally $O(n^4)$.

EMERS computes the set of edit operations that transform the recognized tree into the ground-truth tree. Accordingly, EMERS is not a normalized metric but an edit distance, where if both trees are identical EMERS is equal to zero. The edit distance between trees is a well-defined metric but the representation ambiguity of MathML can mean that correct recognition results are considered errors. In Álvaro et al. (2012b) an experiment using two equivalent ground-truths it was shown that the expression recognition rate, computed as the percentage of expressions with EMERS equal to zero, differed by almost 8% depending on the ground-truth used. A canonical form to represent math expressions in MathML is required in order to avoid this problem. Sain et al. (2010) tried to overcome this problem by converting the MathML to LaTeX and then converting the LaTeX back to MathML.

As with global metrics, the computed error value accounts for the entire expression but the source of the errors is not explicitly known. The set of edit operations is provided and we could compute if they were related to symbols or tags, but segmentation mistakes could not be detected and would become symbols and tags errors.

Finally, the authors propose two options for computing the error: every edit operation has the same cost, or it depends on the baseline (using the concept of *level* defined in previous sections) in which the edit operators are done. The default EMERS value is computed using the weighted version, and this results in a non-symmetrical distance in some cases.

4.3. IMEGE: Image-Based Mathematical Expression Global Error

Although the same math expression can have multiple valid representations, an intuitive idea is that the image generated from those encodings should be the same. That is the main idea of IMEGE, a proposal for computing an error metric using the rendered formulas directly (Álvaro et al., 2013).

Given a recognition result of a certain expression and its ground-truth we want to evaluate the quality of this result. The image representation of a math expression can be generated from its string codification (e.g. LaTeX or MathML). Next we explain the process for computing the recognition error (IMEGE) by using an image-matching model (IDM) and an evaluation algorithm (BIDM).

[1] Available at http://www.isical.ac.in/ utpal/resources.php

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 203 of 404

4.3.1. Image-Matching Model (IDM)

In order to obtain a matching between two images, one option is to compute a two-dimensional warping between them. Keysers et al. (2007) presented several deformation models for image classification, and the Image Distortion Model (IDM) represented the best compromise between computational complexity and evaluation accuracy. Therefore, the IDM was used to perform a two-dimensional matching between two images.

The IDM is a zero-order model of image variability (Keysers et al., 2007). This model uses a mapping function with absolute constraints; hence, it is computationally much simpler than a two-dimensional warping. Its lack of constraints is compensated using a local gradient image context window. This model obtains a dissimilitude measure from one image to another so that if two images are identical, their distance is equal to zero.

The IDM has two parameters: warp range (w) and context window size (c). The algorithm requires each pixel in the test image to be mapped to a pixel within the reference image not more than w pixels from the place it would take in a linear matching. Over all these possible mappings, the best matching pixel is determined using the $c \times c$ local gradient context window by minimizing the difference with the test image pixel. The contribution of both parameters is different for each pixel. The warp range w constrains the set of possible mappings and the $c \times c$ context window computes the difference between the horizontal and vertical derivatives for each mapping. It should be noted that these parameters need to be tuned.

4.3.2. The Evaluation Algorithm (BIDM)

Once we have a model that is able to detect similar regions of two images, we want to use this information to compute an error measure between them. Starting from the IDM-distance algorithm presented in Keysers et al. (2007), we proposed the Binary IDM (BIDM) evaluation algorithm (defined in Algorithm 1). First, instead of calculating the vertical and horizontal derivatives using Sobel filters, these derivatives are computed using the method described in Toselli et al. (2004). Next, the double loop computes the IDM distance for each pixel, and these values are stored individually. Then, the difference between each pixel of the test image and the most similar pixel found in the reference image can be represented as a gray-scale image (Figure 17c-1). At this point, we have a dissimilitude value for each pixel of the test image. However, rather than knowing how different a pixel is, we want to know whether or not a pixel is correct. This is achieved by normalizing the distance values in the range $[0, 255]$ and then performing a binarization process using Otsu's method (Otsu, 1979) (Figure 17c-2). Finally, we intersect the foreground pixels of the test image with the binarized mapping values (like an error mask), and, as a result, we know which pixels are properly recognized and which are incorrectly recognized (Figure 17c-3). Since the background pixels do not provide information, the number of correct pixels is normalized by the foreground pixels.

The time complexity of the algorithm is $O(IJw^2c^2)$, where $I \times J$ are the test image dimensions, w is the warp range parameter, and c is the local gradient context window size. It is important to note that in practice both w and c take low values compared to the image sizes.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 204 of 404

input : test image A ($I \times J$)
reference image B ($X \times Y$)
warp range w
context window size c
output: BIDM(w, c) from A to B
begin
$A^v = \text{vertical\_derivative}(A)$
$A^h = \text{horizontal\_derivative}(A)$
$B^v = \text{vertical\_derivative}(B)$
$B^h = \text{horizontal\_derivative}(B)$

for $i = 1$ *to* I **do**
for $j = 1$ *to* J **do**
$i' = \lfloor i\frac{X}{I} \rfloor$, $j' = \lfloor j\frac{Y}{J} \rfloor$, $z = \lfloor \frac{c}{2} \rfloor$;
$S_1 = \{1, \ldots, X\} \cap \{i' - w, \ldots, i' + w\}$;
$S_2 = \{1, \ldots, Y\} \cap \{j' - w, \ldots, j' + w\}$;

$$\text{map}(i,j) = \min_{\substack{x \in S_1 \\ y \in S_2}} \sum_{m=-z}^{z} \sum_{n=-z}^{z} (A^v_{i+n,j+m} - B^v_{x+n,y+m})^2 + (A^h_{i+n,j+m} - B^h_{x+n,y+m})^2$$

end
end

normalize_depth(map, 255)
binarize(map) *//Otsu's method*

$\text{fg} = \{(x,y) \mid A(x,y) < 255\}$ *//Foreground pixels*
$\text{cp} = \text{fg} \cap \{(x,y) \mid map(x,y) = 0\}$ *//Correct pixels*

return $\frac{|cp|}{|fg|}$ *//Correct pixels ratio*
end

Algorithm 1: Binary IDM (BIDM) evaluation algorithm.

4.3.3. Recognition Error (IMEGE)

The BIDM algorithm computes the number of pixels of a test image that are correctly allocated in another reference image according to the IDM model. The algorithm that we used followed the concepts of precision and recall to compute the Image-based Mathematical Expression Global Error (IMEGE).[2] Firstly, we compute the BIDM value from the test image to the reference (precision p). Secondly, we compute the same value from the reference image to the test image (recall r). Finally, both values are combined using the harmonic

[2]Software available at https://github.com/falvaro/imege

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 205 of 404

mean $f_1 = 2(p \cdot r)/(p + r)$, and we obtain the final error value. Figure 17 illustrates an example of this process.

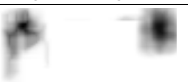
a) Mathematical expression recognition result

ground-truth = {x^2 + 1^3}
recognition = {x2 + 1}

b) Image generation from ground-truth and recognition

$\text{img}_1 =$ $x^2 + 1^3$ $\text{img}_2 =$ $x2 + 1$

c) BIDM computation in both directions

	$\text{img}_2 \rightarrow \text{img}_1$	$\text{img}_1 \rightarrow \text{img}_2$
1		
2		
3	$x2 + 1$	$x^2 + 1^3$
4	precision $= \frac{1489 \text{ ok}}{2197 \text{ fg}} = 0.6777$	recall $= \frac{1429 \text{ ok}}{2338 \text{ fg}} = 0.6112$

d) Recognition global error

$f_1(\text{precision}, \text{recall}) = 0.6427$
$\text{error} = 100(1 - 0.6427) = \mathbf{35.73}$

Figure 17. Example of the procedure for computing the IMEGE measure given a math expression recognition and its ground-truth in LaTeX.

Rendering the image of a math expression encoding copes with the problem of representation ambiguity. IMEGE provides a normalized value in the range $[0, 100]$ than can be interpreted as a visual error (as human beings do) and is not as pessimistic as expression recognition rate. IMEGE can not distinguish the source of the errors although it can identify the misrecognized zones of the math expression. As a visual error, misrecognitions involving larger symbols would affect more pixels than errors produced by smaller symbols. Given that this measure takes the global recognition information into account, it can be very helpful to complement the expression recognition rate and symbol related metrics in order to assess the performance of a system.

4.4. Label Graphs

Zanibbi et al. (2011) proposed a set of performance metrics for on-line handwritten mathematical expressions based on representing the expressions as *label graphs*. A label graph is a directed graph over primitives represented using adjacency matrices. In a label graph, nodes represent primitives, while edges define primitive segmentation (*merge* relationships) and object relationships. Given a math expression, a label graph is constructed from a symbol layout tree (see Figure 7) where the strokes in a symbol are split into separate nodes.

Samsung Exhibit 1014
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 206 of 404

Each stroke keeps the spatial relationship of its associated symbol, and the nodes inherit the spatial relationships of their ancestors in the layout tree.

Figure 18 shows an example of on-line handwritten math expression and two label graphs: a label graph for its ground-truth, and a label graph for a recognition result containing errors. Each label graph is displayed so that the dashed edges show the inherited relationships. The adjacency matrix representation is also provided, where the diagonal of the matrix represents the symbol class of each stroke and other cells provide primitive pair labels. These pairs encode the spatial relationships (right, superscript, etc.), where underscore (_) identifies unlabeled strokes or no-relationship, and an asterisk ($*$) represents two strokes in the same symbol (Zanibbi et al., 2013).

Since the label graph representation contains the information of a mathematical expression at all levels (symbols, segmentation and structure), several metrics can be computed.

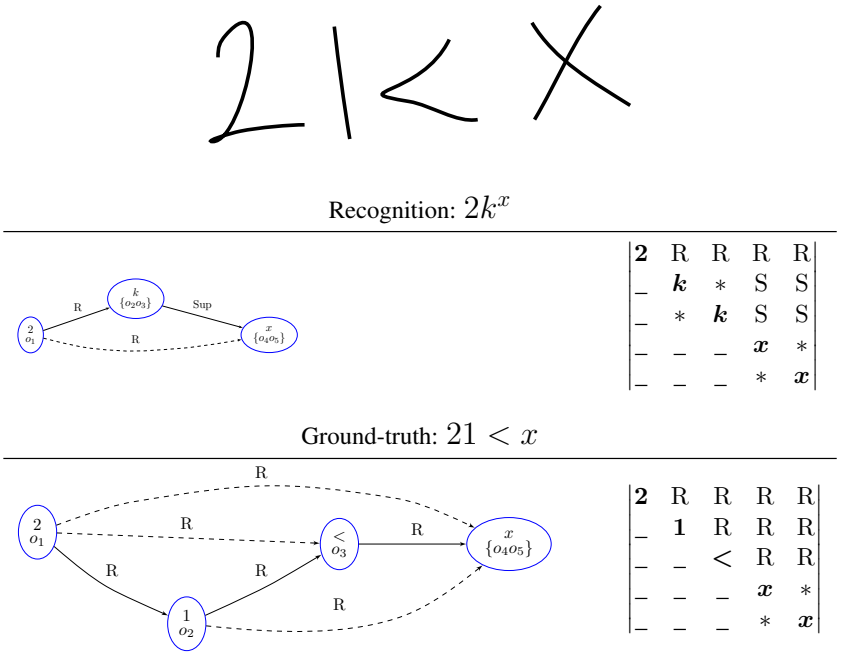


Figure 18. Example of label graph representation of an on-line handwritten math expression recognition and its ground-truth. The dashed edges are inherited relationships.

Given a math expression composed of n strokes, its ground-truth label graph, and the label graph of a recognition result, Zanibbi et al. (2011) defined the following set of metrics. First, metrics for specific errors:

- Classification error (ΔC): the number of strokes that have different symbol classes (elements of the diagonal of the adjacency matrix) in the label graphs.
- Layout error (ΔL): the number of disagreeing edge labels in the label graphs (off-diagonal elements of the adjacency matrix). This error can be decomposed as the sum of segmentation error (ΔS) and relationships error (ΔR), depending on the type

Cerence Operating Company, IPR2024-01267
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 207 of 404

of label of the edges.

Second, metrics at expression level that provide an overall error for a recognition result:

- ΔB_n: the number of disagreeing stroke labels and relationships between two graphs, i.e. the Hamming distance between the matrices of both label graphs. This metric can be computed as
$$\Delta B_n = \frac{\Delta C + \Delta L}{n^2}$$
This metric will result in more distance for layout errors ($n(n-1)$ elements) than for classification errors (n elements) because is not weighted. For this reason, the next metric was also proposed.
- ΔE: the average per-stroke classification, segmentation and layout errors, so that the three types of errors are weighted more equally. It is calculated as
$$\Delta E = \frac{1}{3}\left(\frac{\Delta C}{n} + \sqrt{\frac{\Delta S}{n(n-1)}} + \sqrt{\frac{\Delta L}{n(n-1)}}\right)$$

In the recognition example of Figure 18, we can see that the symbols 1 and $<$ have been incorrectly grouped as a letter k, and the relationship with the letter x has been incorrectly detected as superscript. The error metrics previously described for this example are:

- $\Delta C = 2$; $\quad \{k \rightarrow 1,\ k \rightarrow\ <\}$
- $\Delta S = 2$; $\quad \{* \rightarrow \_\ ,\ * \rightarrow \mathrm{R}\}$
- $\Delta R = 4$; $\quad \{\mathrm{S} \rightarrow \mathrm{R},\ \mathrm{S} \rightarrow \mathrm{R},\ \mathrm{S} \rightarrow \mathrm{R},\ \mathrm{S} \rightarrow \mathrm{R}\}$
- $\Delta L = \Delta S + \Delta R = 6$
- $\Delta B_n = \frac{2+6}{5^2} = \frac{8}{25} = 0.32$
- $\Delta E = \frac{1}{3}\left(\frac{2}{5} + \sqrt{\frac{2}{20}} + \sqrt{\frac{6}{20}}\right) = 0.4213$

All these label graph Hamming distances are proper metrics (Zanibbi et al., 2011), and the time complexity for the expression level metrics is $O(n^2)$, although in practice only the labeled edges have to be compared, which is much faster for sparse graphs. Furthermore, label graph metrics precisely determine the different types of errors at all levels, which is very useful information. Also, the representation ambiguity of formats like LaTeX or MathML is tackled by the label graph representation and the inheritance of relationships.

Finally, precision and recall at object level (symbols) can also be computed from this representation (Zanibbi et al., 2013). The last editions of the CROHME competition use these metrics for assessing the performance of the systems (Mouchère et al., 2013; Mouchère et al., 2014), thus these metrics are becoming the current standard of the state-of-the-art literature in mathematical expression recognition, thanks to the authors that released a great set of open-source tools for computing them[3].

It should be noted that this set of metrics is based on strokes, i.e. for on-line handwritten mathematical expressions. However, the authors pointed out that they can be applied to

[3]Label Graph Evaluation Library at http://www.cs.rit.edu/ dprl/Software.html

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 208 of 404

images using pixels or connected components as primitives, as well as to other related problems like chemical diagrams, flowchart recognition or tables (Zanibbi et al., 2011, 2013). In order to do so, the ground-truth has to be provided at primitive level. In the CROHME competitions, the InkML format contains all the required information to build the label graphs, but using only LaTeX or MathML annotation would not be enough.

4.5. String Matching

One of the most common formats for representing mathematical expressions is LaTeX. Recently, Pavan Kumar et al. (2014) proposed a structure encoded string representation computed from LaTeX, although authors comment that it can be extracted from MathML and any other format. They address the problem of performance evaluation of mathematical expression recognition as computing the Levenshtein edit distance between two strings.

Mathematical expressions are two-dimensional structures, while strings are a one-dimensional format. The proposed string encoding is based on considering symbols left-to-right and two regions for each mathematical symbol, where *top-left*, *above* and *top-right* spatial relationships are encoded in the *northern* region, and *bottom-left*, *below* and *bottom-right* spatial relationships are encoded in the *southern* region. This string encoding deals with representation ambiguity by always processing symbols left-to-right and the northern region before the southern region.

Pavan Kumar et al. (2014) also defined additional symbols to encode the math expressions as strings, like an *empty* symbol in order to handle special cases. A northern region is delimited by start (N_s) and end (N_e) marks, and a southern region has equivalent start (S_s) and end (S_e) marks. For example, any of the LaTeX strings of Figure 16 encoding the math expression $x_a^2 + 1$ would be represented by the string: $x, N_s, 2, N_e, S_s, a, S_e, +, 1$.

This metric follows the same methodology than EMERS with some differences. The string representation makes the edit distance easier, so that time complexity is $O(n^2)$ while EMERS was $O(n^4)$. Also, string representation does not have the problem of considering symbol types as errors like EMERS in MathML. However, computing the string format requires handling several cases and it is not as straightforward as using a MathML tree. Like EMERS, the string matching also reports an edit distance that is not normalized, so that if two strings are identical their distance is zero. The source of the errors (classification, segmentation or structure) is not detected, and the distance could not be symmetric if the error is weighted taking into account the concept of *levels*.

It seems a good alternative to EMERS, handling better the representation ambiguity and having less time complexity. Although label graph metrics are more complete, it could be useful when only the LaTeX or MathML is provided as ground-truth. Unfortunately, at this moment the software for computing these metrics is not publicly available.

5. Experimentation

We developed a system[4] that implements the parsing algorithm of the approach proposed in this chapter[5]. In order to evaluate the performance of this proposal, we carried out several

[4] https://github.com/falvaro/seshat

[5] Demo available at http://cat.prhlt.upv.es/mer

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 209 of 404

experiments using a large public dataset from a recent international competition (Mouchère et al., 2013).

5.1. Dataset

The CROHME 2013 competition (Mouchère et al., 2013) released a large resource for mathematical expression recognition as a result of combining and normalizing several datasets. Currently, it represents the largest public dataset of on-line handwritten mathematical expressions. This dataset has $8,836$ mathematical expressions for training and a test set of 671 math expressions. The number of mathematical symbol classes is 101.

We reported results on the competition test set in order to provide comparable results. The metrics reported are based on label graphs as described in the Label Graphs Section. We needed a validation set for adjusting the models and parameters used in the parsing algorithm. For this reason, we set 500 mathematical expressions apart from the training set, so that both sets had the same distribution of symbols. Therefore, the sets used in the experimentation described in this study were a training set made up of $8,336$ expressions (81K symbols), a validation set with 500 expressions (5K symbols) and the CROHME 2013 test set containing 671 mathematical expressions (6K symbols).

5.2. Estimation of Models

The proposal presented in this study has several probabilistic models that are used as a source of information during the recognition process of a mathematical expression. Below we outline how each model is estimated.

5.2.1. Symbol Classification Model

We trained two BLSTM-RNN using the RNNLIB library[6], one for on-line classification and the other for off-line classification. We used the same configuration for both BLSTM-RNN. The only difference was the size of the input layer, which was determined by the feature set: 7 inputs for on-line data and 9 inputs for off-line data. The output layer size was 101, i.e. the number of symbol classes, and the forward and backward hidden layers each contained 100 LSTM memory blocks.

The network weights were initialized with a Gaussian distribution of mean 0 and standard deviation 0.1. The network was trained using on-line gradient descent with a learning rate of 0.0001 and a momentum of 0.9. This configuration obtained good results in both handwritten text recognition (Graves et al., 2009) and handwritten mathematical symbol classification (Álvaro et al., 2013). We trained each network until the error ceased to improve on the validation set for 50 epochs.

5.2.2. Symbol Segmentation Model

Given the expressions of the training set, we extracted all admissible segmentation hypotheses $b_i \in \mathcal{B}$. Thus, we obtained 360K 4-dimensional samples of symbol segmentations, where 7.5% samples corresponded to proper symbol segmentations and 92.5% samples

[6] http://sourceforge.net/projects/rnnl/

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 210 of 404

were incorrect segmentations. From the validation set we extracted 35K samples: 4.4% correct and 95.6% incorrect segmentations.

We trained the GMMs using the training samples and the Expectation-Maximization algorithm. The parameters of the model were chosen by minimizing the error when classifying the segmentation hypotheses of the validation set. The number of mixtures in the final GMMs was 5.

5.2.3. Symbol Duration Model

To estimate the duration model we used a simple ratio between the number of times a given symbol s_i was composed of l_i strokes and the total number of samples of that symbol found in the training set. The values for a mathematical symbol class were smoothed using add-one smoothing (Jurafsky and Martin, 2008) in the range $[1, L_{\max}]$.

5.2.4. Spatial Relationships Model

We extracted the spatial relationships between symbols and subexpressions for all the mathematical expressions of the training set. This extraction had to be carried out taking into account the centroid computation for each type of symbol and the combination of regions. By doing so, we obtained 68K 9-dimensional samples of spatial relationships for training and 4K samples for validation. The distribution of the classes was quite unbalanced. In the training set the *right* relationships were very common (71.84%), while *inside* and *mroot* relationships were infrequent (2.51% and 0.19%). The percentage of samples of *below*, *subscript* and *superscript* were about 12.08%, 5.56% and 7.82%, respectively.

Once we had the set of training samples, we trained a GMM per class using the Expectation-Maximization algorithm. The number of mixtures of the final GMMs was 5, since it represented the best trade-off between the complexity of the model and the classification error in the validation set.

5.2.5. 2D-PCFG Estimation

The structure of the mathematical expressions is described by a 2D-PCFG. Since the rules of mathematical notation are well-known, starting from the CFG provided by the organizers of the CROHME competition, we manually modified it to improve the modeling of some structures. We also added productions that increase ambiguity in order to model certain structures like the relations between categories of symbols (uppercase/lowercase letters, numbers, etc.). However, the probabilities of the productions of the 2D-PCFG have to be estimated.

A usual approach to estimating probabilistic grammars is the Viterbi score (Ney, 1992). We recognized the expressions of the training set in order to obtain the most likely derivation trees according to the grammar. As recognizing the training set will introduce errors into the computed trees, we used constrained parsing (Álvaro et al., 2012a) to obtain the parse tree that best represents each training sample. Figure 19 shows an scheme of this training process. Thus, the probability of a production $A \rightarrow \alpha$ was calculated as

$$p(A \rightarrow \alpha) = \frac{c(A \rightarrow \alpha)}{c(A)}$$

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 211 of 404

where $c(A \rightarrow \alpha)$ is the number of times that the rule $A \rightarrow \alpha$ was used when recognizing the training set, and $c(A)$ is the total number of productions used that have A as left-hand operator. In order to account for unseen events, we smoothed the probabilities using add-one smoothing (Jurafsky and Martin, 2008).

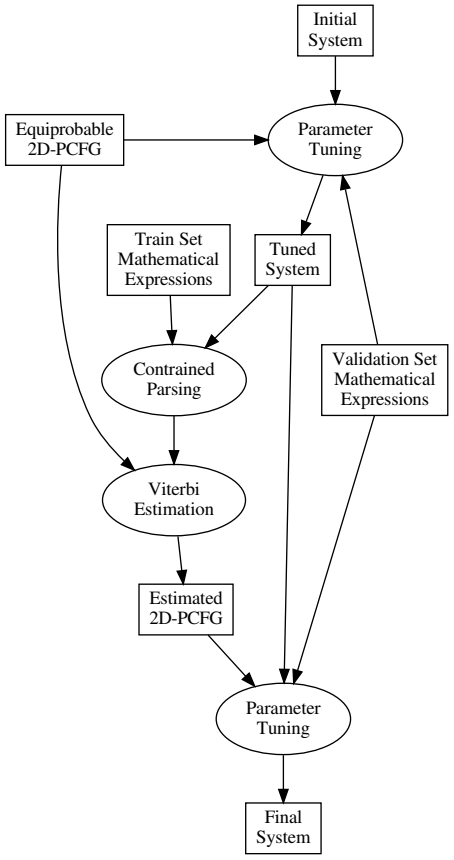


Figure 19. Diagram of the process for training the final mathematical expression recognition system.

5.3. Parameter Setting

The parsing algorithm defined in the Parsing Algorithm Section has two steps. First, when the parsing table is initialized, multi-stroke symbol hypotheses are introduced as subproblems of $1 \leq n \leq L_{\max}$ strokes following Equation (10). Then, in the general case, the parsing table is filled with hypotheses of $n \geq 2$ strokes by combining smaller subproblems using Equation (11). This produces a scaling problem of the probabilities.

The probability of a hypothesis of size $n \geq 2$ created in the initialization step is the product of four probabilities, while the hypothesis resulting from the combination of subproblems will involve $6n - 2$ terms in the calculation. The different probabilistic distributions have been estimated separately, which leads to values in different scales.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 212 of 404

For these reasons, following (Luo et al., 2008), we assigned different exponential weights to each model probability, and we also added an insertion penalty in the initialization step (Equation (10)). These parameters alleviate the scaling differences of the probabilities. The weights help to adjust the contribution of each model to the final probability, since some sources of information are more relevant than the others.

The parameters of the parsing algorithm are: insertion penalty, exponential weights, segmentation distance threshold and maximum number of strokes per symbol ($L_{\max}$). We set $L_{\max} = 4$ because it accounts for 99.81% of the symbols in the dataset. The remaining parameters were set initially to 1.0 and we tuned them using the downhill simplex algorithm (Nelder and Mead, 1965) minimizing the ΔE metric (Zanibbi et al., 2011) when recognizing the validation set (Figure 19).

5.4. Experiments and Results

After training all the models and parameters using the training set, we used the system that obtained the best results on the validation set to classify the test set of the CROHME 2013 competition. Eight systems participated, including a preliminary version of this model (system IV). All but two of the systems used only the competition training set ($8,836$ mathematical expressions). System III also used 468 additional mathematical expressions, and System VII was trained using roughly $30,000$ mathematical expressions from a private corpus. The description of each system can be found in (Mouchère et al., 2013).

Table 1 shows the performance metrics at symbol level, and Table 2 shows results at stroke level. Results show that system VII performed the best, obtaining very good re-

Table 1. Object-level evaluation for the CROHME'13 dataset. Systems sorted by decreasing recall for correct symbol segmentation and classification (Seg+Class).

	Segments		Seg+Class		DAG Relations	
	Rec.	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**
VII	97.9	98.1	93.0	93.3	95.2	95.5
seshat	**92.0**	**90.7**	**82.2**	**81.0**	**88.0**	**82.0**
IV*	85.0	87.1	73.9	75.8	76.3	79.9
VIII	90.3	86.9	73.8	71.0	73.0	77.7
V	84.5	86.5	66.7	68.3	72.6	74.3
II	80.7	86.4	66.4	71.1	45.8	63.0
III	85.2	77.9	62.6	57.3	88.5	78.3
VI	57.9	47.3	47.7	39.0	31.8	70.0
I	46.9	38.4	25.2	20.6	33.7	71.6

sults. It was declared the strongest system in the competition. However, as they used a large private dataset we were not able to fairly compare its performance to that of the other systems. System IV was a preliminary version of `seshat` and was declared the best system trained using only the CROHME training dataset. The main differences between System IV and `seshat` are as follows: `seshat` includes off-line information in the symbol classifier; symbol segmentation and spatial relationships classification are carried out

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 213 of 404

by GMMs in seshat and by SVMs in system IV; seshat uses *visibility* between strokes and a clustering-based penalty; and the probabilities of the grammar were not estimated in System IV.

Table 2. Stroke-level evaluation for the CROHME'13 dataset. Systems sorted by increasing ΔE. ΔB_n and ΔE are measured on directed labels graphs (8548 strokes; 81007 (undirected) stroke pairs).

	Label Hamming Distances				μ error (%)	
	Strokes	**Pairs**	**Seg**	**Rel**	ΔB_n	ΔE
VII	537	1 777	170	1 607	2.4	4.3
seshat	**1 583**	**7 829**	**700**	**7 129**	**6.9**	**13.2**
IV*	2 187	9 493	1 201	8 292	10.1	18.3
VIII	2 302	15 644	4 945	10 699	12.1	19.3
II	2 748	19 768	1 527	18 241	13.9	22.0
V	2 898	10 803	1 228	9 575	12.7	22.8
III	3 415	15 135	1 262	13 873	15.0	26.2
VI	4 768	43 893	5 094	38 799	27.6	36.7
I	6 543	41 295	5 849	35 446	26.8	41.6

The system presented in this study significantly outperformed the other systems that were trained using the CROHME dataset at all levels. At symbol level, symbol classification of correctly segmented symbols of seshat obtained recall and precision of about 82.2% and 81.0%, while the next best system (system VIII) obtained 73.8 and 71.0%. The absolute difference was more than 8% in recall and 10% in precision. Regarding spatial relationships, recall and precision for seshat stood at 88% and 82%, while System VIII gave 73% and 77.7%. This translates into an absolute difference of about 14% in recall and 4.3% in precision. Results at stroke-level were also better than those of the systems trained only using the CROHME dataset. The systems were ranked according to the global error metric ΔE, where seshat had 13.2%, some 6.1% less than next best system (19.3). Table 3 shows that the confidence interval (Bisani and Ney, 2004) was $13.2\% \pm 0.9$ at 95% of confidence. We were not able to obtain confidence interval for the other systems because the final system outputs were not freely available.

In addition to the experimentation comparing our proposed model to other systems, it is interesting to see how each of the stochastic sources contribute to overall system performance. For this reason we also carried out an experiment to observe this behavior. Some models are mandatory for recognizing mathematical expressions: the symbol classifier, the spatial relationships classifier and the grammar. We performed an experiment using only these models (base system), then added the remaining models one by one. Also, the grammar initially had equiprobable productions and then we compared the performance when the probabilities of the rules were estimated.

Table 3 shows the changes in system performance when each source of information is added. Global error metrics ΔB_n and ΔE consistently decreased with each added feature. Confidence intervals computed for ΔE showed that the improvements were significant

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 214 of 404

Table 3. Contribution to overall system performance of the different sources of information used. The models and features listed in each row are cumulative, so that the system shown in the last row includes all information sources. Confidence intervals (Bisani and Ney, 2004) were computed for ΔE.

	Segments		Seg+Class		Relations		Label Hamming Distances				μ error (%)	
	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Strokes	Pairs	Seg	Rel	ΔB_n	ΔE
Base system	87.6	83.6	78.3	74.7	87.0	75.0	1 987	11 999	1 172	10 827	9.1	16.5 ± 1.0
+ dur. mod.	88.0	84.2	78.7	75.3	86.2	75.5	1 928	12 055	1 138	10 917	9.0	15.8 ± 1.1
+ seg. mod.	90.9	90.4	81.3	80.8	81.5	73.7	1 634	10 272	834	9 438	7.6	13.7 ± 1.0
+ rel. pen.	91.8	91.3	82.0	81.5	81.4	74.0	1 556	10 397	724	9 673	7.5	13.2 ± 0.9
+ gram. est.	92.0	90.7	82.2	81.0	88.0	82.0	1 583	7 829	700	7 129	6.9	13.2 ± 0.9

from the first row to the last row. Symbol segmentation and symbol recognition also improved with each addition. It is interesting that, when no segmentation model was used, symbol segmentation still gave good results. This was the case because the parameters of the integrated approach converged to high values of the insertion penalty and low values of the segmentation distance threshold. In this way, the parameters of the system itself could alleviate the lack of a segmentation model. In any case, when the segmentation model was included, system performance improved significantly. Furthermore, we would like to underline that, when the relations penalty was included in the model, the number of hypotheses explored was reduced by 56.7%.

The structural analysis is harder to evaluate. Prior to estimating the grammar probabilities, the results at object level seem to worsen when the segmentation model was included, although at stroke-level the errors in spatial relationships decreased from about $11,000$ to $9,500$ stroke pairs. Because of the inheritance of spatial relations in label graphs (Zanibbi et al., 2011) some types of structural errors can produce more stroke-pair errors than others (Zanibbi et al., 2013). Specifically, when the segmentation model was used, the segmentation distance threshold was approximately twice as high as the value of the threshold in the base system. This had two effects. First, that the system was able to account for more symbol segmentations, as shown by the corresponding metrics. Second, that a bad decision in symbol segmentation can lead to worse structural errors. Nevertheless, the estimation of the probabilities of the grammar led to great improvements in the detection of the structure of the expression with barely any changes in symbol recognition performance.

References

Álvaro, F., Sánchez, J. A., and Bened, J. M. (2012a). Unbiased evaluation of handwritten mathematical expression recognition. In *International Conference on Frontiers in Handwriting Recognition*, pages 181–186.

Álvaro, F., Sánchez, J. A., and Benedí, J. M. (2013). Classification of On-line Mathematical Symbols with Hybrid Features and Recurrent Neural Networks. In *International Conference on Document Analysis and Recognition*, pages 1012–1016.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 215 of 404

Álvaro, F., Sánchez, J.-A., and Benedí, J.-M. (2013). An image-based measure for evaluation of mathematical expression recognition. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 682–690. Springer Berlin Heidelberg.

Álvaro, F., Sánchez, J. A., and Benedí, J. M. (2014). Offline Features for Classifying Handwritten Math Symbols with Recurrent Neural Networks. In *International Conference on Pattern Recognition*, pages 2944–2949.

Álvaro, F., Sánchez, J.-A., and Benedí, J.-M. (2014). Recognition of on-line handwritten mathematical expressions using 2d stochastic context-free grammars and hidden markov models. *Pattern Recognition Letters*, 35(0):58 – 67.

Álvaro, F., Sánchez, J.-A., and Benedí, J.-M. (2016). An integrated grammar-based approach for mathematical expression recognition. *Pattern Recognition*, 51:135 – 147.

Álvaro, F., Snchez, J.-A., and Bened, J.-M. (2012b). Unbiased evaluation of handwritten mathematical expression recognition. In *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 181–186.

Álvaro, F. and Zanibbi, R. (2013). A Shape-Based Layout Descriptor for Classifying Spatial Relationships in Handwritten Math. In *ACM Symposium on Document Engineering*, pages 123–126.

Anderson, R. H. (1967). Syntax-directed recognition of hand-printed two-dimensional mathematics. In *ACM Symposium on Interactive Systems for Experimental Applied Mathematics*, pages 436–459.

Awal, A.-M., Mouchere, H., and Viard-Gaudin, C. (2009). Towards handwritten mathematical expression recognition. In *International Conference on Document Analysis and Recognition*, pages 1046–1050.

Awal, A.-M., Mouchere, H., and Viard-Gaudin, C. (2010). The problem of handwritten mathematical expression recognition evaluation. In *International Conference on Frontiers in Handwriting Recognition*, pages 646–651.

Awal, A.-M., Mouchre, H., and Viard-Gaudin, C. (2014). A global learning approach for an online handwritten mathematical expression recognition system. *Pattern Recognition Letters*, 35(0):68 – 77.

Bisani, M. and Ney, H. (2004). Bootstrap estimates for confidence intervals in ASR performance evaluation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 409–412, Montreal.

Chan, K.-F. and Yeung, D.-Y. (2000). Mathematical expression recognition: a survey. *International Journal on Document Analysis and Recognition*, 3:3–15.

Chan, K.-F. and Yeung, D.-Y. (2001). Error detection, error correction and performance evaluation in on-line mathematical expression recognition. *Pattern Recognition*, 34:1671 – 1684.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 216 of 404

Chou, P. A. (1989). Recognition of equations using a two-dimensional stochastic context-free grammar. In *Visual Communications and Image Processing IV*, volume 1199, pages 852–863.

Eto, Y. and Suzuki, M. (2001). Mathematical formula recognition using virtual link network. In *International Conference on Document Analysis and Recognition*, pages 762–767.

Faure, C. and Wang, Z. (1990). Automatic perception of the structure of handwritten mathematical expressions. In *Computer Processing of Handwriting*, pages 337–361.

Garain, U. and Chaudhuri, B. (2004). Recognition of online handwritten mathematical expressions. *IEEE Trans. on Systems, Man, and Cybernetics - Part B: Cybernetics*, 34(6):2366–2376.

Garain, U. and Chaudhuri, B. (2005). A corpus for ocr research on mathematical expressions. *International Journal on Document Analysis and Recognition*, 7:241–259.

Goodman, J. (1999). Semiring parsing. *Computational Linguistics*, 25(4):573–605.

Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., and Schmidhuber, J. (2009). A novel connectionist system for unconstrained handwriting recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(5):855–868.

Ha, J., Haralick, R., and Phillips, I. (1995). Understanding mathematical expressions from document images. In *International Conference on Document Analysis and Recognition*, volume 2, pages 956–959.

Hu, L. and Zanibbi, R. (2011). HMM-Based Recognition of Online Handwritten Mathematical Symbols Using Segmental K-Means Initialization and a Modified Pen-Up/Down Feature. In *International Conference on Document Analysis and Recognition*, pages 457–462.

Hu, L. and Zanibbi, R. (2013). Segmenting Handwritten Math Symbols Using AdaBoost and Multi-Scale Shape Context Features. In *International Conference on Document Analysis and Recognition*, pages 1180–1184.

Jurafsky, D. and Martin, J. H. (2008). *Speech and Language Processing (2nd edition) (Prentice Hall Series in Artificial Intelligence)*.

Keshari, B. and Watt, S. (2007). Hybrid mathematical symbol recognition using support vector machines. In *International Conference on Document Analysis and Recognition*, volume 2, pages 859 –863.

Keysers, D., Deselaers, T., Gollan, C., and Ney, H. (2007). Deformation models for image recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(8):1422–1435.

Lapointe, A. and Blostein, D. (2009). Issues in performance evaluation: A case study of math recognition. In *International Conference on Document Analysis and Recognition*, pages 1355–1359.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 217 of 404

Lavirotte, S. and Pottier, L. (1998). Mathematical formula recognition using graph grammar. In *Proceedings of the SPIE*, volume 3305, pages 44–52.

Lehmberg, S., Winkler, H.-J., and Lang, M. (1996). A soft-decision approach for symbol segmentation within handwritten mathematical expressions. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 3434–3437.

Luo, Z., Shi, Y., and Soong, F. (2008). Symbol graph based discriminative training and rescoring for improved math symbol recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1953–1956.

MacLean, S. and Labahn, G. (2010). Elastic matching in linear time and constant space. *IAPR Internatioanl Workshop on Document Analysis Systems*, pages 551–554.

MacLean, S. and Labahn, G. (2013). A new approach for recognizing handwritten mathematics using relational grammars and fuzzy sets. *International Journal on Document Analysis and Recognition*, 16(2):139–163.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.

Marti, U.-V. and Bunke, H. (2001). Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(01):65–90.

Mouchère, H., Viard-Gaudin, C., Zanibbi, R., and Garain, U. (2014). ICFHR 2014 Competition on Recognition of On-Line Handwritten Mathematical Expressions (CROHME 2014). In *Frontiers in Handwriting Recognition (ICFHR), International Conference on*, pages 791–796.

Mouchère, H., Viard-Gaudin, C., Zanibbi, R., Garain, U., and Kim, D. H. (2013). ICDAR 2013 CROHME: Third International Competition on Recognition of Online Handwritten Mathematical Expressions. In *International Conference on Document Analysis and Recognition*, pages 1428–1432.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, 7:308–313.

Ney, H. (1992). Stochastic Grammars and Pattern Recognition. In *Speech Recognition and Understanding*, volume 75, pages 319–344.

Okamoto, M., Imai, H., and Takagi, K. (2001). Performance evaluation of a robust method for mathematical expression recognition. In *Proc. 6th International Conference on Document Analysis and Recognition (ICDAR'01)*, pages 121–128.

Okamoto, N. and B., M. (1991). Recognition of mathematical expressions by using the layout structures of symbols. In *International Conference on Document Analysis and Recognition*, pages 242–250.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 218 of 404

Otsu, N. (1979). A Threshold Selection Method from Gray-level Histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66.

Pavan Kumar, P., Agarwal, A., and Bhagvati, C. (2014). A string matching based algorithm for performance evaluation of mathematical expression recognition. *Sadhana*, 39(1):63–79.

Pearlmutter, B. (1989). Learning state space trajectories in recurrent neural networks. In *International Joint Conference on Neural Networks*, volume 2, pages 365–372.

Pra, D. and Hlav, V. (2007). Mathematical formulae recognition using 2d grammars. *International Conference on Document Analysis and Recognition*, 2:849–853.

Sain, K., Dasgupta, A., and Garain, U. (2010). EMERS: a tree matching based performance evaluation of mathematical expression recognition systems. *IJDAR*, pages 1–11.

Schuster, M. and Paliwal, K. (1997). Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45(11):2673–2681.

Shi, Y., Li, H., and Soong, F. K. (2007). A Unified Framework for Symbol Segmentation and Recognition of Handwritten Mathematical Expressions. In *International Conference on Document Analysis and Recognition*, pages 854–858.

Shi, Y. and Soong, F. (2008). A symbol graph based handwritten math expression recognition. In *International Conference on Pattern Recognition*, pages 1–4.

Sibson, R. (1973). SLINK: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1):30–34.

Tapia, E. and Rojas, R. (2004). Recognition of on-line handwritten mathematical expressions using a minimum spanning tree construction and symbol dominance. In *Graphics Recognition. Recent Advances and Perspectives*, volume 3088 of *Lecture Notes in Computer Science*, pages 329–340. Springer Berlin Heidelberg.

Thammano, A. and Rugkunchon, S. (2006). A neural network model for online handwritten mathematical symbol recognition. In *Intelligent Computing*, volume 4113, pages 292–298.

Toselli, A., Juan, A., and Vidal, E. (2004). Spontaneous handwriting recognition and classification. In *Proc. of the 17th International Conference on Pattern Recognition*, pages 433–436, Cambridge, UK.

Toselli, A., Pastor, M., and Vidal, E. (2007). On-line handwriting recognition system for tamil handwritten characters. In *Pattern Recognition and Image Analysis*, volume 4477, pages 370–377.

Winkler, H.-J. (1996). HMM-based handwritten symbol recognition using on-line and off-line features. In *IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 3438–3441.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 219 of 404

Yamamoto, R., Sako, S., Nishimoto, T., and Sagayama, S. (2006). On-line recognition of handwritten mathematical expressions based on stroke-based stochastic context-free grammar. *IEIC Technical Report*.

Zanibbi, R. and Blostein, D. (2012). Recognition and retrieval of mathematical expressions. *International Journal on Document Analysis and Recognition*, 15(4):331–357.

Zanibbi, R., Blostein, D., and Cordy, J. (2002). Recognizing mathematical expressions using tree transformation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(11):1–13.

Zanibbi, R., Mouchère, H., and Viard-Gaudin, C. (2013). Evaluating structural pattern recognition for handwritten math via primitive label graphs. In *Document Recognition and Retrieval (DRR)*.

Zanibbi, R., Pillay, A., Mouchère, H., Viard-Gaudin, C., and Blostein, D. (2011). Stroke-based performance metrics for handwritten mathematical expressions. In *International Conference on Document Analysis and Recognition*, pages 334–338.

Zhang, L., Blostein, D., and Zanibbi, R. (2005). Using fuzzy logic to analyze superscript and subscript relations in handwritten mathematical expressions. In *International Conference on Document Analysis and Recognition*, volume 2, pages 972–976.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 220 of 404

In: Handwriting: Recognition, Development and Analysis ISBN: 978-1-53611-937-4
Editors: Byron L. D. Bezerra et al. © 2017 Nova Science Publishers, Inc.

Chapter 8

ONLINE HANDWRITING RECOGNITION OF INDIAN SCRIPTS

Umapada Pal[1,*] ***and Nilanjana Bhattacharya***[2,†]
[1]Computer Vision and Pattern Recognition Unit
Indian Statistical Institute, Kolkata, India
[2]Bose Institute, Kolkata, India

1. Introduction

Handwriting recognition of unconstrained Indian text is difficult because of the presence of many complex shaped compound characters as well as variability involved in the writing style of different individuals. Though there are many published works on Latin, Chinese or Japanese handwriting recognition, recognition of Indian scripts is yet to be extensively investigated and reliable recognition system with high accuracy is yet to be obtained. A number of studies (Pal et al., 2012; Jayadevan et al., 2011) have been done for offline recognition of Indian scripts like Bangla, Devanagari, Gurmukhi, Tamil, Telugu, Oriya, etc. Only a few works have been reported on online recognition of cursive Indian scripts. In the following Sub-section, we present the state of the art of online handwriting recognition studies of Indian scripts.

1.1. State of the Art Online Handwriting Recognition of Indian Scripts

Many pieces of work are available in the literature for online recognition of different scripts (Plamondon and Srihari, 2000). Some of the recent works on feature extraction and online recognition are reviewed here. The task of the features is to provide a systematic representation of the unstructured input data. Features in general should be discriminative in order to facilitate the classification and as low in numbers as possible to avoid the "curse of dimensionality" (Bellman, 1957). In handwriting recognition domain, no clear standard has

*E-mail address: umapada@isical.ac.in.
†E-mail address: nilibht@gmail.com.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

emerged, and usually each research group uses their own set of, mostly hand-crafted, features. This is quite remarkable since in related field, such as speech recognition, a standard set of features, e.g. the Mel-frequency cepstral coefficients (MFCC), is widely accepted (Greenberg et al., 2004).

Some works are available on online isolated Bangla character recognition in Roy et al. (2007) and Mondal et al. (2010). A few works are available on online Bangla cursive text recognition. In Bhattacharya U. et al. (2008), online handwritten words were segmented estimating the position of the headline of the word. Preprocessing operations such as smoothing and re-sampling of points were done before feature extraction. They used 77 features considering 9 chain-code directions. Modified quadratic discriminant function (MQDF) classifier was used for recognition. They did not consider any word involving compound character. In Bhattacharya and Pal (2012), A system for segmentation and recognition of Bangla online handwritten text was described. Authors identified 85 stroke classes used in cursive Bangla handwriting. Cursive words were explicitly segmented into primitives using an automatic segmentation algorithm. Segmentation procedure made use of some rules discovered analyzing different joining patterns of Bangla characters. Next, the primitives obtained from the input word were recognized using 64 chain code histogram features and SVM classifier (Vapnik, 1995; Burges, 1998). A similar method for segmentation and recognition of Bangla online handwritten text containing compound characters was described in Bhattacharya N. et al. (2013). In Fink et al. (2010), the authors divided each stroke of the pre-processed word sample into several sub-strokes using the angle incurred while writing. Features used were length, angle, width, normalized x and y coordinates. Then HMM was used for recognition. A work on holistic word recognition can be found in Samanta et al. (2014). Authors tried to explicitly segment the word such that segmentation points represent approximate character boundaries. They used 24 angles with respect to neighboring points as features along with some of the NPen++ features reported in Jaeger et al. (2001). HMM-based classifier was used considering both forward and reverse sequences of features from handwritten words.

In Swethalakshmi (2007), representations of strokes based on spatio-temporal, spectral and spatio-structural features were proposed for Devanagari and Tamil scripts. Studies on stroke classification were performed using support vector machines for the three proposed representations. A rule-based approach was proposed for recognition of characters from recognized strokes. Though this work did not handle cursive text, the performance of a hidden. Markov model-based classifier was compared for the case of spatiotemporal representation of strokes. Spatiotemporal feature vector contained the x and y co-ordinates of the online points. The work in Bharath and Madhvanath (2012) addressed lexicon-free and lexicon-driven recognition problems for Type-1 (discretely written), Type-2 (cursively written) and Type-3 (written with delayed strokes) words of Devanagari and Tamil scripts. They used the features introduced in NPen++ recognizer (Jaeger et al., 2001) and implemented HMM-based classifier for recognition.

Naz et al. (2014) surveyed the optical character recognition (OCR) and online character recognition literature with reference to the Urdu-like cursive scripts. In particular, Urdu, Pushto, and Sindhi languages are discussed, with the emphasis being on the Nasta'liq and Naskh scripts. The works are analyzed comparing with a typical OCR pipeline with an emphasis on the pre-processing, segmentation, feature extraction, classification, and recog-

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 222 of 404

nition.

Different systems for handwriting recognition use different features to represent the input text. Even after decades of research, no favourable decision on a best-practice exists and many features are carefully hand-crafted. To facilitate the design phase for on-line handwriting systems, Frinken et al. (2014b) proposed an unsupervised feature generation approach based on dissimilarity space embedding (DSE) of local neighbourhoods around the points along the trajectory. DSE has a high capability of discriminative representation and hence beneficial for classification (Pekalska and Duin, 2005). They compared the approach with a state-of-the-art feature extraction method and demonstrated its superiority. The fundamental idea of the feature extraction process, inspired from previously mentioned works on DSE (Fischer et al., 2010; Pekalska and Duin, 2005; Riesen and Bunke, 2009), was to first create a fixed set of reference points then use the list of dissimilarities of the current element to the reference points as features. To find the reference prototypes, two unsupervised approaches were investigated - clustering based and random prototype selection. While a wide variety of prototype selection strategies are known, these two approaches show a robust performance over different types of datasets (Bunke and Riesen, 2012). Bidirectional long short-term memory (BLSTM) neural network (Graves et al., 2009) was used for Bangla words recognition.

We see that the most successful approaches to classifying temporal pattern involve hidden Markov Models (Bishop, 1992; Sun and Jelinek, 1999; Rabiner, 1989) or recurrent neural networks. HMM has been proved to be successful to classify temporal patterns (Cho et al., 1995); while long-term dependencies within sequences have been successfully addressed using BLSTM neural networks for handwriting recognition of Latin script (Graves et al., 2009). In Frinken et al. (2014a), authors investigated different encoding schemes of Bangla compound characters and compared the recognition accuracies. They proposed to model complex characters not as unique symbols, represented by individual nodes in the output layer. Instead, they exploited the BLSTM NN property of long-distance-dependent classification. Only basic strokes were classified and special nodes were used to react to semantic changes in the writing, i.e., distinguishing inter-character spaces from intra-character spaces. This was the first work, to the knowledge of the authors, which explored the applicability of BLSTM neural network to Bangla words containing the compound character.

Texts containing crossing out, repeated writing of the same stroke several times, and over-writing are very common in practical life. These three types of writing (crossing out, repetition and over-writing) can be called as “noise”. A practical problem in data acquisition and recognition is the detection and removal of noise from handwriting. The only work on detection of noisy regions in online words was done in Bhattacharya et al. (2015). Authors claimed that the method was able to detect noise from both online and offline words. Different density-based features were proposed to distinguish between “relevant” and “unwanted” (or noisy) parts of writing and a 2-class HMM based offline classifier was used for classification into clean and noisy parts in a word.

In this chapter, we choose to describe an approach for recognition of Bangla online words. Bangla is an important script and it is very difficult to recognize.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 223 of 404

2. Properties of Bangla

Bangla is the second most popular language in India and the fifth most popular language in the world. More than 200 million people speak in Bangla and Bangla script is used in Assamese and Manipuri languages in addition to Bangla language. The set of basic characters of Bangla consists of 11 vowels and 39 consonants. As a result, there are 50 different shapes in the Bangla basic character set. The concept of upper/lower case is absent in this script. Figure 1 shows ideal (printed) forms of these 50 basic character shapes.

In Bangla, a vowel (except for the first vowel) can take modified form and we call it a vowel modifier. Ideal (printed) shapes of these vowel modifiers corresponding to 10 vowels with a basic character KA are shown in Figure 2. Similarly, consonants can also take modified form. Figure 3 shows consonant modifiers with a basic character BA.

A consonant or a vowel following a consonant sometimes takes a compound orthographic shape, which we call compound character. Compound characters can be combinations of two or three characters. Modifiers can be used with compound characters which may result in more complicated shapes. Examples of some Bangla compound character formations are shown in 4-a. Occasionally, the combination of two basic characters forms a new shape as shown in the first two rows of Figure 4-a. In the third and fourth rows of 4-a one of the constituent characters of the compound character retains its shape and the other constituent character reduces its size in the compound character. In the compound character shown in the fifth row, two characters sit side by side in compounding, but the size of the first character is slightly reduced. In the compound characters depicted in the sixth and seventh rows are formed by three basic characters where the shape of none of its constituent basic characters can be found. Since the formation of compound characters is different and people write these in many ways, it is very difficult to recognize Bangla compound characters.

Main difficulty of any character recognition system is the shape similarity. It can be noted that because of handwritten style, two different characters in Bangla may look very close. For example, in Figure 4-b, some similar shaped Bangla compound character pairs are shown. Such shape similarity makes the recognition system more complex.

Unconstrained Bangla handwriting is usually cursive. In one stroke, the writer can write a part of a character or one or more characters. It is found that a single stroke may contain up to 6 characters and modifiers. Also in Bangla, the most of the touchings of characters in a word occur in the region of word's headline or sirorekha portion (Pal et al., 2012) in contrast to English handwriting where the touchings occur in the lower part of the word shape. For characters forming a compound character, joining occurs at the end part of the first character (which may be in upper or middle or lower region of the word) with the beginning of the next character.

On the other hand, several single characters are written in a variety of ways - in a single stroke or in more than one stroke. From the statistical analysis, it is found that the minimum number of stroke used to write a Bangla character is 1 and the maximum number is 6. Hence online recognition of Bangla is a difficult task.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 224 of 404



Figure 1. Bangla basic characters above (vowels are in green, consonants in brown) and their respective codes for future reference.



Figure 2. Vowel modifiers of Bangla and their respective codes with basic character KA.

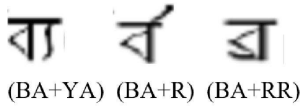


Figure 3. Consonant modifiers of Bangla and their respective codes with basic character BA.

3. An Approach for Stroke Segmentation and Recognition of Bangla Online Words

In this chapter, we describe the approach proposed in Bhattacharya N. et al. (2013) for stroke segmentation and recognition from Bangla online handwritten text containing both basic and compound characters and all types of modifiers. The algorithm is robust against

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 225 of 404

Constituent basic characters		Compound characters
ক + র	=	ক্র
ঙ + গ	=	ঙ্গ
জ + ব	=	জ্ব
ন + ট	=	ন্ট
চ + ছ	=	চ্ছ
ক + ষ + ম	=	ক্ষ্ম
স + ত + র	=	স্ত্র

(a)



(b)

Figure 4. (a) Bangla compound character formation from basic characters. (b) Similar shaped compound characters.

various types of stroke connections as well as shape variations. For segmentation of strokes, some rules are discovered analyzing different joining patterns of Bangla characters. By manually analyzing different strokes after segmentation, 251 distinct primitive classes are obtained. Directional features of 64 dimensions are extracted to recognize the segmented primitives using SVM classifier.

3.1. Data Collection and Ground Truth Generation

There is no available dataset on online Bangla cursive text. For the experiments of segmentation and recognition of Bangla strokes, a set of 4984 Bangla words written by 100 writers were collected using Wacom tablet. No restriction was imposed on writing. Writers were requested to write words from a given a lexicon of 170 words. These words were selected in such a way that all basic and compound characters and all types of modifiers were present in the total set. Input data consisted of (x, y) coordinates along the trajectory of the pen together with positions of pen-downs (stroke starting points).

A text file was built with ground truths of segmentation for all input word files. Each row of this file contained input filename, the number of ideal segmentation points and their x, y co-ordinates. For each input file, output segmentation points are to be compared with ground-truth-file and accuracy is automatically calculated without manual intervention. Similarly, ground truth file was created for automatic recognition accuracy calculation. Each row of this file contained input filename, primitive ids of segmented primitives.

3.2. Segmentation of Cursive Word

Segmentation is one of the important phases of handwriting recognition in which data are represented at character or stroke level so that nature of each character or stroke can be studied individually. Joining occur mostly in the upper portion of the word in case of Bangla words. And also, Bangla writing goes from left to right. Considering these facts automatic segmentation algorithm was designed.

Suppose, middle zone (or busy zone) is defined as the region between two lines-

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 226 of 404

TOP_LINE and BOTTOM_LINE. For segmentation purpose, up and down zones are defined, as depicted in Figure 5(a). From topmost row of the word to (TOP_LINE + t1) row is up zone and (TOP_LINE + t2) row to down most row of the word is down zone. Here, t1=height of busy zone/3. t2= height of busy zone/2. The height of busy zone= BOTTOM_LINE - TOP_LINE.

Online points are described as up, down or don't know points according to their belonging to up zone, down zone or outside these zones. If the pen tip goes from down zone to up zone and then again come to down zone, two characters or modifiers may be touching in the up zone and hence the stroke should be segmented (Figure 5 (b)). Because of this, for each stroke, stroke movement patterns like "down-> up->down" are found, i.e. "any number of down points followed by any number of up points followed by any number of down points" within the stroke (don't points are simply ignored). For such pattern, segmentation is done at the highest point of up zone of the touching. Such segmentation point is called as candidate segmentation point. For "down->up->down" stroke, from the first "down", find down most point. From second "down" also find the down most point. Find the point which is higher (nearer to up points) among these two down most points. Call it "HIGHER_DOWN".

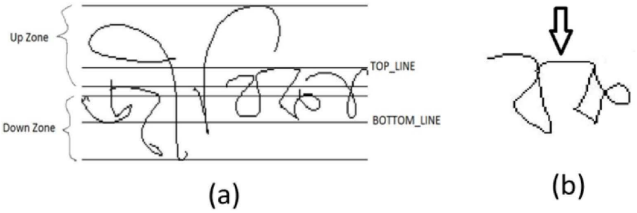


Figure 5. (a) TOP_LINE, BOTTOM_LINE, up zone and down zone in a word. (b) Touching of BA and KA (stroke movement form: up->down->up->down->up).

Now, the candidate points are validated to avoid over-segmentation. Using positional information and stroke patterns, two levels of validations are performed as follows:

(I) VALIDATION OF CANDIDATE POINTS AT LEVEL-1: Candidate points are found within some characters where "down->up->down" pattern is present. These are not joining points. Level-1 validation is done to find only valid joining points. The position of the candidate point is to be tested with respect to the position of HIGHER_DOWN, BOTTOM_LINE of the busy zone, and also with respect to stroke height. The following four conditions are tested:

1. $r(HIGHER\_DOWN) - r(candidatepoint) > (heightofbusyzone \times 40\%)$
2. $r(HIGHER\_DOWN) - r(candidatepoint) > (heightofthestroke \times 30\%)$
3. $r(BOTTOM\_LINE) - r(candidatepoint) > (heightofbusyzone \times 60\%)$
4. $r(downmostpointofthestroke) - r(candidatepoint) > (heightofthestroke \times 40\%)$

where r(x) means row value of x.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 227 of 404

If all of these 4 conditions are satisfied by a candidate segmentation point, it is a valid segmentation point.

(II) VALIDATION OF CANDIDATE POINTS AT LEVEL-2:

Some rules are implemented which are discovered by analyzing stroke patterns of Bangla writing. The observations are as follows:

As Bangla writing goes from left to right, the end point of a stroke consisting of more than one character is always at the right side of the start point. If the stroke consists of only a character or a part of a character this relationship between the start point and end point does not always hold. Hence, the segmentation rules are as follows:

a. End point of a connected stroke should be at the right side of start point of the stroke, i.e. *c(end point)* $>$ *c(start point)*, where *c(x)* means column value of *x*. Otherwise, candidate segmentation point is cancelled.

b. End point of a connected stroke should be at the right side of previous validated segmentation point of the stroke, i.e. *c(end point)* $>$ *c(previous segmentation point)*. Otherwise, candidate segmentation point is cancelled.

Examples of some of the results obtained before and after Level-2 validation are shown in Figure 6. Different strokes of input word are depicted in different colors and the segmentation points are shown in red on the strokes.

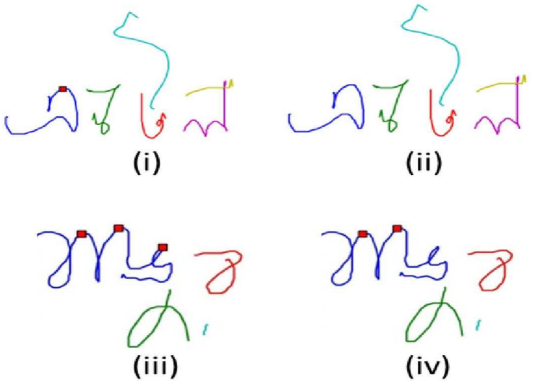


Figure 6. Candidate segmentation points are shown by small solid red squares. (i) Before applying Rule-(a): E is over-segmented. (ii) After applying Rule-(a). (iii) Before applying Rule-(b): NGA is over-segmented. (iv) After applying Rule-(b).

3.3. Stroke Analysis

At first, a general analysis is done on Bangla alphabet to find the number of stroke classes which are sufficient to cover all characters and modifiers. If parts of different characters look similar, they are assigned with a single stroke-id. On the other hand, stroke classes representing one particular character differ from writer to writer. For example, Figure 7 (ii)

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 228 of 404

and Figure 7 (iv) shows two GAs written by different writers. The left stroke of first GA (Figure 7 (ii)) is similar to the right stroke of KA (Figure 7 (i)). Also, the left stroke of second GA (Figure reffig:7 (iv)) is similar to the left stroke in SA (Figure 7 (iii)). Hence, in the ground truth file, their codes are also considered similar.

Next, the stroke classes are analyzed with respect to the segmentation algorithm. There are 11 additional stroke classes because of over-segmentation. If all types of joining between characters and modifiers are considered, it is found that some characters can be joined with vowel modifiers like U, UU, R and consonant modifiers like R, RR within a single stroke. As these modifiers can not be segmented from characters, these joined strokes are considered as separate stroke classes. Thus 30 additional primitives are obtained for GA+UU, DA+U etc. Some new shapes are obtained for the combination of character and modifier (for example, HA+U, BHA+RR etc).

Now we come to compound characters. As we have mentioned in the proposed segmentation approach, in case of compound character, if the first character ends at its right side and in the upper region of the word, the compound character will got segmented by the algorithm. Some compound characters can not be segmented because the joining occurred in the lower part of the first character. These compounds are considered to be new classes. Occasionally, constituent characters of the compound character form a new shape. For example, HA+MA, KA+SSA etc. There are 11 such compounds which are new classes. Some additional classes are also obtained for joining of compound characters with modifiers. For 3-character compounds, segmentation may occur differently depending on the length of each of the three characters. All the possibilities of segmentation are considered to get all possible primitive classes. Finally, considering all the above cases, a set of 251 distinct primitive classes is found. Table 1 shows a few examples of primitive classes and the characters in which these primitives are used.



Figure 7. (i) KA, (ii) GA, (iii) SA, (iv) GA. Right (black) stroke of KA and left (black) stroke of GA in (ii) are the same. Left (green) stroke of SA and left (black) stroke of GA in (iv) are the same.

3.4. Feature and Classifier

Here, the 64-dimensional feature vector is used for high-speed primitive recognition. Each primitive is divided into 4x4 cells, i.e. 16 cells and frequencies of the direction codes are computed in these cells. Chain code of four directions [0 (horizontal), 1 (+45 degrees from positive x-axis), 2 (vertical) and 3 (+135 degrees from positive x-axis)] are only used. Figure 8 illustrates chain code directions. It is assumed that chain code of direction 0 and 4, 1 and 5, 2 and 6, 3 and 7, are equivalent features because it is found that strokes of characters BA, LA, PA, GA and modifiers E, II, AU, R (consonant) can be written with

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 229 of 404

Table 1. Primitives and their respective characters

Primitive	Characters in which the primitive is used
	O, AU, NYA, GA+U, SHA+U, SSA+NNA, TA+TA, KA+TA, JA+NYA, NYA+CA
	NA+MA, NA+DDA, NA+DA, NA+TTA, NA+TTHA, PA+NA, GA+NA
	E, AI, NYA, KA+RR, TA+RR
	NGA, RA+U, DA+RR+U, BHA+RR+U
	DDA, RRA, U, UU, JA, NGA, JA+NYA

different orders of pen points within the stroke making the directions just opposite. Thus, for each cell, we get four integer values representing the histograms of the four direction codes. So, 16x4=64 features are found for each primitive. These features are normalized by dividing by the maximum value.

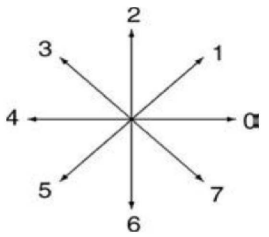


Figure 8. Chain code directions for feature computation.

In this experiment, a Support Vector Machine (SVM) classifier is used for primitive recognition. The SVM is originally defined for two-class problems and it looks for the optimal hyper plane which maximizes the distance and the margin, between the nearest examples of both classes, named support vectors (SVs). Given a training database of M data: $\{x_m \mid m = 1...M\}$, the linear SVM classifier is then defined as:

$$f(x) = \sum_j a_j x_j.x + b$$

where$\{x_j\}$ is the set of support vectors and the parameters α_j and b have been determined by solving the quadratic problem (Vapnik, 1995). The linear SVM can be extended to various non-linear variants (Vapnik, 1995). In this experiment, the Gaussian kernel SVM outperformed other non-linear SVM kernels.

A total number of 27,344 primitive samples are obtained after segmentation. 50% of these samples are used for training and rest for testing. Word is recognized using a table look-up approach by matching the sequence of primitives. If the exact entry is not found, the nearest entry is considered.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 230 of 404

4. Results and Discussion

4.1. Segmentation Result

The ground truth file is used to verify accuracy of automatic segmentation algorithm. From the dataset of 4984 words, the segmentation scheme gives an accuracy of 97.89% which is very encouraging. Figure 9 shows some examples of correctly segmented strokes while Figure 10 shows examples of incorrectly segmented strokes.

4.2. Segmentation Error Analysis

Here let us analyze why the segmentation errors occur. We can see in Figure 10 (i) that modifier AA is not segmented because its height is small (which should not happen) and it does not reach the down zone. In Figure 10 (ii), modifier I is not segmented because it does not reach the down zone. On the other hand, in Figure 10 (iii), character NA is over-segmented because it reaches from down zone to up zone and then it comes to down zone. This part of NA should not reach the up zone in an ideal case. Similarly, in Figure 10 (iv), character CHA is over-segmented as it reaches the up zone.

4.3. Recognition Result

We can see the recognition results for words containing only basic characters and modifiers, for words containing at least one compound character, and the combined result in row (A), row (B) and row (C) of Table 2, respectively. From the combined experiment on 13,672 test samples, 97.45% primitive recognition accuracy is obtained where the sample set of 251 primitive classes includes basic characters/compound characters/modifiers, parts of basic /compound characters/ modifiers having meaningful structural information, and parts incurred while joining.

Table 2. Primitive Recognition Result

Dataset	Average Primitive Recognition rate
(A) Words containing only basic characters and modifiers	97.68%
(B) Words containing basic and compound characters and modifiers	96.35%
(C) Combined dataset	97.45%

4.4. Primitive Recognition Error Analysis

Now, let us discuss the causes of primitive recognition errors. Characters GHA, YY, THA, KHA, PHA (Shown in Figure 1) look very similar and hence generate some misclassifications. Similarly, characters CA and DDHA; compound characters NA+TA and NA+DDA; HA+MA and KA+SSA; GA+NA and GA+LA generate some errors because of their similarity. In summary, the cause of errors is the shape similarity among the primitives.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 231 of 404

4.5. Results obtained from other works

Here, we report some of the other published results. In Mondal et al. (2010), authors reported basic character recognition accuracy of 81.55 % (using point-float feature in HMM) to 91.01% (using chain-code feature in Nearest Neighbour classifier) on 8,616 test character samples, where samples include only 50 basic characters. In Bhattacharya U. et al. (2008), authors selected a lexicon of 100 Bangla words and reported that 3.1% of the segmented strokes suffered from under segmentation. Only properly segmented strokes were used for training and testing of the classifier. Recognition error obtained was 1.22% at stroke level considering 73 stroke classes. In Fink et al. (2010), authors reported recognition accuracy of 88% (for holistic recognition - which treats all word samples separately) to 93.1% (for context-dependent sub-word units recognition) on 6,516 test word samples where samples include 50 Indian city names.

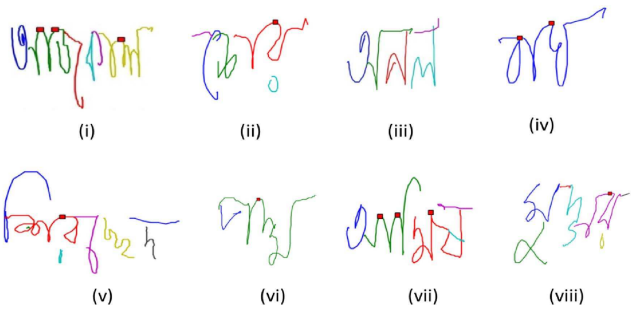


Figure 9. Examples of words which are correctly segmented.

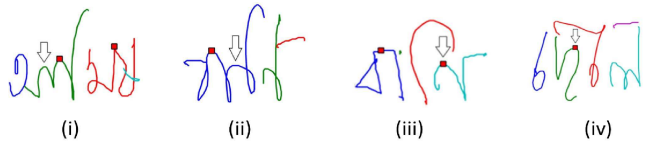


Figure 10. Examples of words which are not segmented correctly (first two words are under segmented, next two words are over segmented). Arrows indicate the positions where under-segmentations and over- segmentations have occurred.

Conclusion

Both segmentation, as well as recognition of online Indian scripts, is yet to get full attention from researchers. Because of complex nature of character formation as well as the presence of many complex shaped compound characters, handwriting recognition of Indian script is very challenging. This chapter discusses the state of the art of online handwriting recognition of main Indian scripts and also presents a work for rigorous primitive analysis and recognition taking into account both Bangla (Bengali) basic and compound characters. We

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 232 of 404

noted that the number of character classes in Bangla is more than the number of exhaustive primitive classes in Bangla. At first, a rule-based scheme is used to segment online handwritten Bangla cursive words into primitives. Using directional features in SVM classifier, primitives are recognized. Word is recognized from the sequence of primitives. Finally, results obtained from the method as well as other published results are discussed and causes of errors are studied.

References

Bishop C. (1992). *Pattern Recognition & Machine Learning*. Elsevier BV.

Bellman, R. E. (1957). *Dynamic Programming*. Princeton University Press.

Bharath, A. and Madhvanath, S. (2012). HMM-Based Lexicon-Driven and Lexicon-Free Word Recognition for Online Handwritten Indic Scripts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):670–682.

Bhattacharya, N., Frinken, V., Pal, U., and Roy, P. P. (2015). Overwriting repetition and crossing-out detection in online handwritten text. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 680–684. IEEE.

Bhattacharya, N. and Pal, U. (2012). Stroke segmentation and recognition from bangla online handwritten text. In *2012 International Conference on Frontiers in Handwriting Recognition*, pages 736–741. IEEE.

Bhattacharya, N., Pal, U., and Kimura, F. (2013). A system for bangla online handwritten text. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1367–1371. IEEE.

Bhattacharya, U., Nigam, A., Rawat, Y. S., and K., P. S. (2008). An analytic scheme for online handwritten bangla cursive word recognition. In *Proceedings of the 2008 10th International Conference on Frontiers in Handwriting Recognition*, ICFHR '08, pages 320–325.

Bunke, H. and Riesen, K. (2012). Towards the unification of structural and statistical pattern recognition. *Pattern Recognition Letters*, 33(7):811–825.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Cho, W., Lee, S.-W., and Kim, J. H. (1995). Modeling and recognition of cursive words with hidden markov models. *Pattern Recognition*, 28(12):1941–1953.

Fink, G. A., Vajda, S., Bhattacharya, U., Parui, S. K., and Chaudhuri, B. B. (2010). Online bangla word recognition using sub-stroke level features and hidden markov models. In *2010 12th International Conference on Frontiers in Handwriting Recognition*, pages 393–398. IEEE.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 233 of 404

Fischer, A., Riesen, K., and Bunke, H. (2010). Graph similarity features for HMM-based handwriting recognition in historical documents. In *2010 12th International Conference on Frontiers in Handwriting Recognition*, pages 253–258. IEEE.

Frinken, V., Bhattacharya, N., and Pal, U. (2014). Design of unsupervised feature extraction system for on-line bangla handwriting recognition. In *2014 11th IAPR International Workshop on Document Analysis Systems*, pages 355–359. IEEE.

Frinken V., Bhattacharya N., Uchida S., and Pal U. (2014). Improved BLSTM Neural Networks for Recognition of On-Line Bangla Complex Words. In *IAPR Joint International Workshops on Statistical Techniques in Pattern Recognition + Structural and Syntactic Pattern Recognition*. Lecture Notes in Computer Science, pages 404–413. Springer.

Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., and Schmidhuber, J. (2009). A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868.

Greenberg, S., Popper, A. N., Ainsworth, W. A., and Fay, R. R. (2004). *Speech Processing in the Auditory System*. Springer-Verlag New York Inc.

Swethalakshmi, S. (2007). *Online Handwritten Character Recognition for Devanagari and Tamil Scripts Using Support Vector Machines*. PhD thesis, Indian Insttitute of Technology.

Jaeger, S., Manke, S., Reichert, J., and Waibel, A. (2001). Online handwriting recognition: the NPen++ recognizer. *International Journal on Document Analysis and Recognition*, 3(3):169–180.

Jayadevan, R., Kolhe, S. R., Patil, P. M., and Pal, U. (2011). Offline recognition of devanagari script: A survey. *Trans. Sys. Man Cyber Part C*, 41(6):782–796.

Mondal, T., Bhattacharya, U., Parui, S. K., Das, K., and Mandalapu, D. (2010). On-line handwriting recognition of indian scripts - the first benchmark. In *Proceedings of the 2010 12th International Conference on Frontiers in Handwriting Recognition*, ICFHR '10, pages 200–205, Washington, DC, USA. IEEE.

Naz, S., Hayat, K., Razzak, M. I., Anwar, M. W., Madani, S. A., and Khan, S. U. (2014). The optical character recognition of urdu-like cursive scripts. *Pattern Recognition*, 47(3):1229–1248.

Pal, U., Jayadevan, R., and Sharma, N. (2012). Handwriting recognition in indian regional scripts: A survey of offline techniques. 11(1):1–35.

Pekalska, E. and Duin, R. P. W. (2005). *The Dissimilarity Representation for Pattern Recognition - Foundations and Applications*. World Scientific Publishing Co. Pte. Ltd.

Plamondon, R. and Srihari, S. N. (2000). On-line and off-line handwriting recognition: A comprehensive survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(1):63–84.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 234 of 404

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, 77(2):257–286.

Riesen, K. and Bunke, H. (2009). Graph classification based on vector space embedding. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(06):1053–1081.

Roy, K., Sharma, N., Pal, T., and Pal, U. (2007). Online bangla handwriting recognition system. In *International Conference on Advances in Pattern Recognition*, pages 121–126.

Samanta, O., Bhattacharya, U., and Parui, S. (2014). Smoothing of HMM parameters for efficient recognition of online handwriting. *Pattern Recognition*, 47(11):3614–3629.

Sun, D. X. and Jelinek, F. (1999). Statistical methods for speech recognition. *Journal of the American Statistical Association*, 94(446):650.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer Nature.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 235 of 404

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

In: Handwriting: Recognition, Development and Analysis ISBN: 978-1-53611-937-4
Editors: Byron L. D. Bezerra et al. © 2017 Nova Science Publishers, Inc.

Chapter 9

HISTORICAL HANDWRITTEN DOCUMENT ANALYSIS OF SOUTHEAST ASIAN PALM LEAF MANUSCRIPTS

***Made Windu Antara Kesiman*[1,*], *Jean-Christophe Burie*[1], *Jean-Marc Ogier*[1], *Gusti Ngurah Made Agus Wibawantara*[2] *and I Made Gede Sunarya*[2]**
[1]Laboratoire Informatique Image Interaction (L3i),
University of La Rochelle, La Rochelle, France
[2]Laboratory of Cultural Informatics (LCI),
University of Pendidikan Ganesha, Singaraja, Bali, Indonesia

1. Introduction

Ancient manuscripts record many pieces of important knowledge about world civilization histories. In Southeast Asia, most of the ancient manuscripts are written on palm leaf. Ancient palm leaf manuscripts are one of the very valuable cultural heritages that store various forms of knowledge and historical records of social life in Southeast Asia. Many palm leaf manuscripts contain information on important issues such as medicines and village regulations that are used as daily guidance. It attracts the historians, philologists, and archaeologists to discover more about the ancient ways of life. The existence of ancient palm leaf manuscripts in Southeast Asia is very important both in term of quantity and variety of historical contents.

For example in Bali, Indonesia, the island's literary works were mostly recorded on dried and treated palm leaves (Figure 1). The dried and treated palm leaf manuscripts in Bali are called lontar. Lontar is written on a dried palm leaf by using some sort small knife. Lontars are inscribed with a special tool called pengerupak. It is made of iron, with its tip sharpened in a triangular shape so it can make both thick and thin inscriptions. The manuscripts were then scrubbed by a natural dyes to leave a black color on the scratched part as text (Figure 2). The writings were incised in one (and/or both) sides of the leaf and

*E-mail address: made_windu_antara.kesiman@univ-lr.fr (Corresponding author).

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

the script is then blackened with soot. The leaves are held and linked together by a string that passes through the central holes and knotted at the outer ends.

The Balinese palm leaf manuscripts were written in the Balinese script and the Balinese language, in the ancient literary texts composed in the old Javanese language of Kawi and Sanskrit. The epic of lontar varies from ordinary texts to Bali's most sacred writings (Figure 3). Many of those epics based on the famous Indian epics of Ramayana and Mahabharata. They include texts on religion, holy formulae, rituals, family genealogies, law codes, treaties on medicine (usadha), arts and architecture, calendars, prose, poems and even magics.

But unfortunately, in reality, the majority of Balinese has never read any lontar because of language obstacles as well as tradition which perceived them as a sacrilege. There is only a limited access to the content of the manuscripts, because of the linguistic difficulties and the fragility of the documents. Balinese script is considered to be one of the complex scripts from Southeast Asia. The alphabet of Balinese script is composed of ± 100 character classes including consonants, vowels, diacritics, and some other special compound characters. In Balinese manuscripts, there is no space between words in a text line. Some characters are written on upper baseline or under the baseline of text line.



Figure 1. Palm tree (left), the dried and treated palm leaves (right).

The physical condition of natural materials from palm leaves certainly cannot fight against time. Usually, palm leaf manuscripts are of poor quality since the documents have degraded over time due to storage conditions. Many discovered lontars are now part of collections of museums and private families. They are in a state of disrepair due to age and due to inadequate storage conditions. Equipment that can be used to protect the palm leaf to prevent rapid deterioration are still relatively few in number. Therefore, the digitization and indexing projects for palm leaf manuscripts were proposed (Kesiman et al., 2015a,b, 2016b; Burie et al., 2016; Kesiman et al., 2016a,c, 2017).

In the last five years, the collection of palm leaf manuscripts in Southeast Asia attracted the attention of researchers in document image analysis. For example, a digitization project for palm leaf manuscripts from Indonesia (Kesiman et al., 2015a,b, 2016b; Burie

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 238 of 404

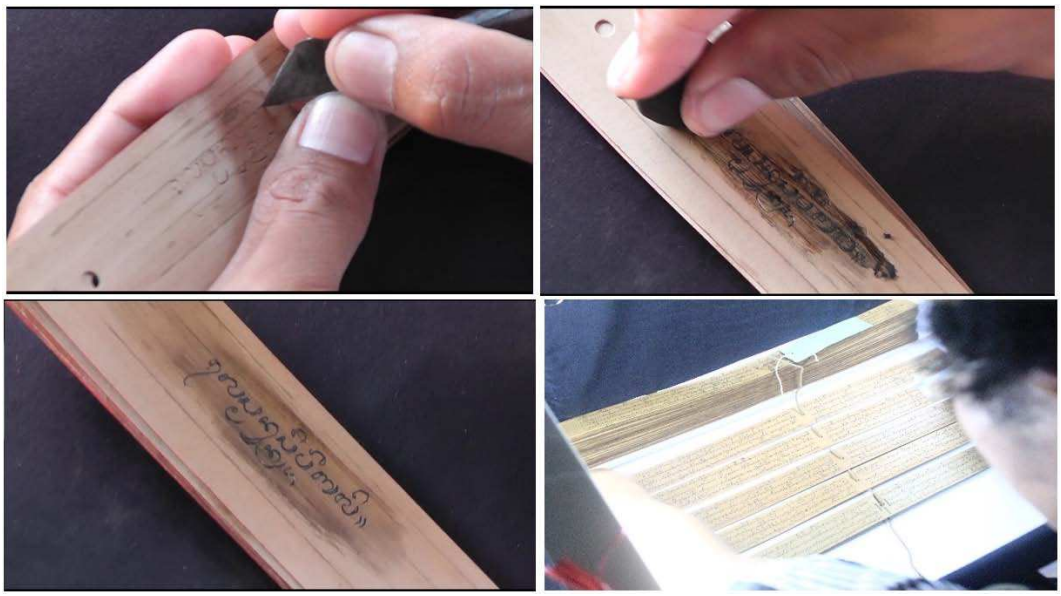


Figure 2. Writing script in lontar with pengerupak.



Figure 3. Balinese palm leaf manuscripts.

et al., 2016; Kesiman et al., 2016a,c, 2017) under the scheme of the AMADI (Ancient Manuscripts Digitization and Indexation) Project, Cambodia[1] and Thailand (Chamchong et al., 2010; Fung and Chamchong, 2010). The AMADI Project works not only to digitize the palm leaf manuscripts, but also to develop an automatic analysis, transcription and indexing system for the manuscripts. Our objectives are to bring added value to digitized palm leaf manuscripts by developing tools to analyze, index and access quickly and efficiently to the content of palm leaf manuscripts, and to make palm leaf manuscripts more accessible, readable and understandable to a wider audience and to scholars and students all over the world. Nowadays, due to the specific characteristics of the physical support of the manuscript, the development of document analysis methods for palm leaf manuscripts in order to extract relevant information is considered as a new research problem in handwritten document analysis (Kesiman et al., 2015a,b, 2016b; Burie et al., 2016; Kesiman et al., 2016a,c, 2017; Chamchong et al., 2010; Chamchong and Fung, 2011, 2012). It ranges wide from binarization process (Kesiman et al., 2015a,b; Burie et al., 2016), text line segmentation (Kesiman et al., 2017), character and text recognition tasks (Burie et al., 2016; Kesiman et al., 2016c) to the word spotting methods.

[1] http://www.khmermanuscripts.org/.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 239 of 404

Ancient palm leaf manuscripts contain artefacts due to aging, foxing, yellowing, marks of strain, local shading effects, with low-intensity variations or poor contrast, random noises, discoloured part, fading (Figure 4). Written on a dried palm leaf by using a sharp pen (which looks like a small knife) and colored with natural dyes, it is hard to separate the text from the background in the binarization process.

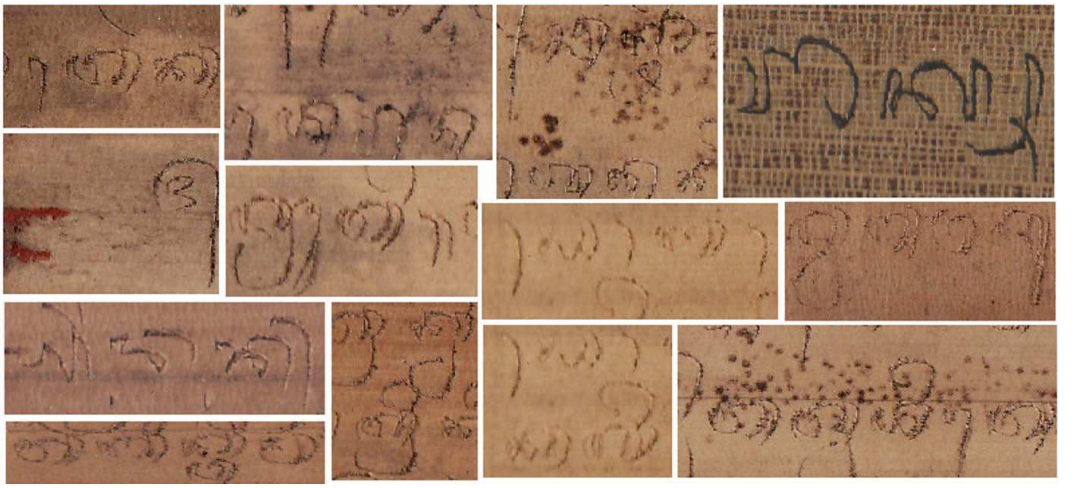


Figure 4. The degradations on palm leaf manuscripts (Kesiman et al., 2015a).

In the OCR task and development, severals deformations in the character shapes are visible due to the merges and fractures of the use of nonstandard fonts. The similarities of distinct character shapes, the overlaps, and interconnection of the neighboring characters further complicate the problem of OCR system (Arica and Yarman-Vural, 2002) (Figure 5). One of the main problems faced when dealing with segmented handwritten character recognition is the ambiguity and illegibility of the characters (Blumenstein et al., 2003). These characteristics provide a suitable condition to test and evaluate the robustness of feature extraction methods which were already proposed for character recognition. Using a character recognition system will help to transcript these ancient documents and translate them to the current language, to give an access to the important information and knowledge in palm leaf manuscript. OCR system is one of the most demanding systems which has to be developed for the collection of palm leaf manuscript images.

This chapter is organized as follow: the following section gives a description about the binarization of palm leaf manuscript images, the construction of ground truth binarized images, and the analysis of ground truth binarized image variability. The section ”Isolated Character Recognition” presents some most commonly used feature extraction methods and describes our proposed combination of features for the isolated character recognition. A segmentation free and training free word spotting method for our palm leaf manuscript images is presented in section ”Word Spotting”. The palm leaf manuscript image dataset used in our experiments and the experimental results are presented respectively in section ”Corpus and Dataset” and ”Experiments”. Conclusions with some prospects for the future works are given in the last section.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 240 of 404

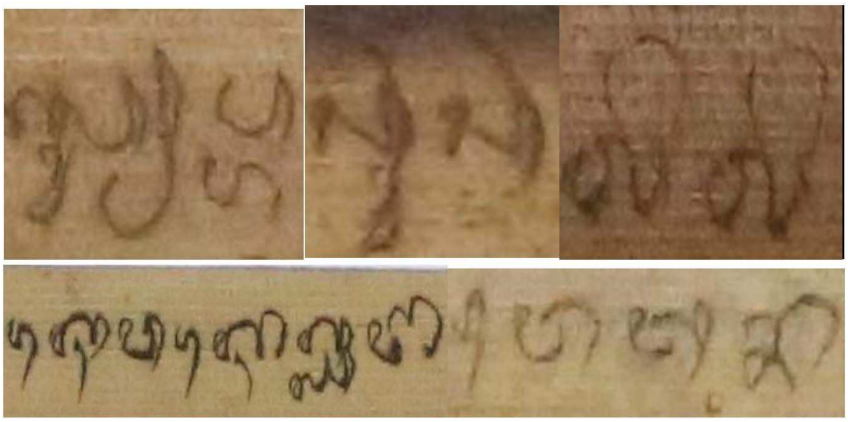


Figure 5. Balinese script on palm leaf manuscripts (Kesiman et al., 2016a).

2. Binarization and Construction of Ground Truth Binarized Images

2.1. The Binarization of Palm Leaf Manuscript Images

With the aim of finding an optimal binarization method for palm leaf manuscript images, some binarization methods which have already been proposed and widely used in document image research community have to be tested and evaluated. We experimented and compared several alternative well-known binarization algorithms on our palm leaf manuscript images. Figure 6 shows the binarized images when applying different method such as Otsu (Pratikakis et al., 2013; Messaoud et al., 2011), Niblack (Khurshid et al., 2009; Rais et al., 2004; Gupta et al., 2007; He et al., 2005; Feng and Tan, 2004), Sauvola (Sauvola and Pietikäinen, 2000), Wolf (Khurshid et al., 2009; Rais et al., 2004), Rais (Rais et al., 2004), NICK (Khurshid et al., 2009), and Howe (Howe, 2013). Since there is no existing ground truth binarized image for our palm leaf manuscripts, we cannot objectively evaluate these results. Therefore, a visual observation process has been applied to compare the results. It is clear that those binarization methods do not give a good binarized image for palm leaf manuscript images. All methods extract unrecognizable characters on palm leaf manuscripts with noise. Therefore, to binarize the images of palm leaf manuscripts, a specific and adapted binarization technique is required.

2.2. The Construction of Ground Truth Binarized Images

To evaluate the performance of binarization methods, two approaches are widely used. The first approach evaluates the binarization methods based on the character recognition rate reached by an OCR system applied on those binarized images (Ntirogiannis et al., 2013). But this approach has been criticized that the binarization method is evaluated in their interaction with other process on document analysis pipeline. The second approach evaluates the binarization methods by comparing pixel-by-pixel the difference between binarized image and a ground truthed binarized image (Pratikakis et al., 2013; Gatos et al., 2011). In the case where the OCR system for some specific Southeast Asian alphabets is not avail-

Chinese Academy of Sciences, Ex. 2014
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 241 of 404

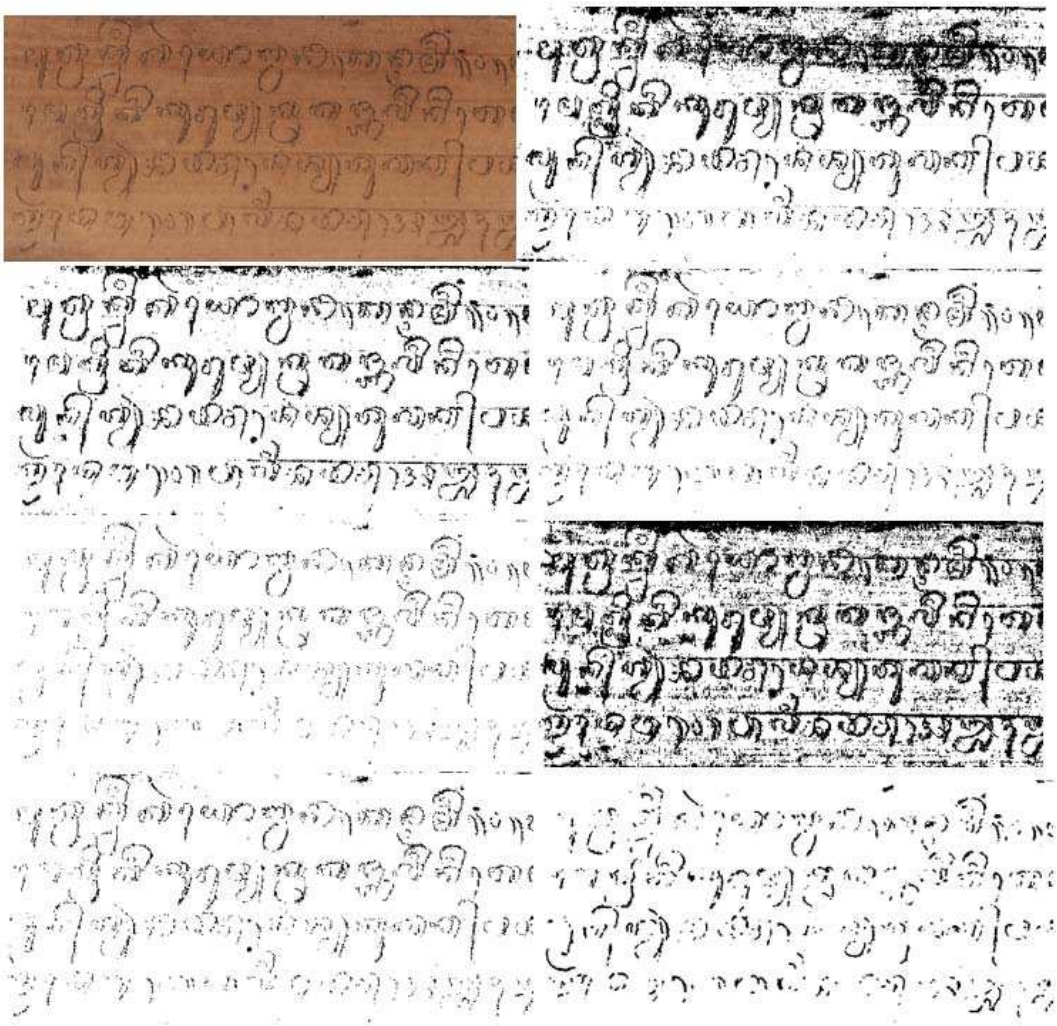


Figure 6. Original image (upper left), and binarized images (up to bottom, left to right) (Kesiman et al., 2015a) using methods of Otsu (Pratikakis et al., 2013; Messaoud et al., 2011), Niblack (Khurshid et al., 2009; Rais et al., 2004; Gupta et al., 2007; He et al., 2005; Feng and Tan, 2004), Sauvola (Sauvola and Pietikäinen, 2000), Wolf (Khurshid et al., 2009; Rais et al., 2004), Rais (Rais et al., 2004), NICK (Khurshid et al., 2009), and Howe (Howe, 2013).

able yet, the ground truth binarized image of palm leaf manuscripts has to be created to be able to quantitatively measure and compare the performance of all binarization methods. Therefore, in order to evaluate and to select an optimal binarization method, creating a new ground truth binarized image of palm leaf manuscripts is a necessary step (Kesiman et al., 2015a).

Manual creation of the ground truth binarized images (e.g. with PixLabeler application (Saund et al., 2009)) is a time-consuming task. Therefore, several semi-automatic frameworks for the construction of ground truth binarized images have been presented

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 242 of 404

(Ntirogiannis et al., 2013, 2008; Nafchi et al., 2013; Bal et al., 2008) to reduce the time of ground truthing process. The human intervention is required only for some necessary but limited tasks. The previous works on construction of ground truth binarized images were especially based on the method used for DIBCO competition series (Pratikakis et al., 2013; Gatos et al., 2011). The need for a specific scheme which adapts and performs better in constructing the ground truth of binarized images for palm leaf manuscripts should be analyzed to achieve a better ground truth for low quality palm leaf manuscripts.

For the DIBCO competition series (Pratikakis et al., 2013), the ground truth binarized images are constructed using a semi-automatic procedure described in (Ntirogiannis et al., 2013). This procedure is adapted and improved by some other works on the construction of ground truth binarized images. For instance, in (Messaoud et al., 2011), a similar method is used to create ground truth of a large document database. In (Nafchi et al., 2013), in order to save user time in manual modification process by an expert, two features of phase congruency are used to pre-process Persian heritage images to generate a rough initial binarized image. In (Bal et al., 2008), the ground truth binarized image of the machine-printed document is constructed by segmenting and clustering the characters during the foreground enhancement step. The user can manually add and remove character model assignments to degraded character instances. Unfortunately, it is impossible to validate a ground truth construction methodology to create a perfect ground truth image from a real image. The ground truth images are normally accepted based on visual observation.

The construction of ground truth binarized images proposed in (Ntirogiannis et al., 2008), consists of several steps: initial binarization process, skeletonization of the characters, manual correction of skeleton, and second skeletonization after manual correction process. The estimated ground truth image is then constructed by dilating the corrected skeleton image, constrained by the character edges (detected using Canny algorithm (Canny, 1986)) and the binarized image under evaluation. The skeleton is dilated until half of the Canny edges intersect each binarized component. The detailed algorithm in pseudo code can be found in (Ntirogiannis et al., 2008). In this method, poor quality of the initial binarized image will directly affect the result of the estimated ground truth. The ground truth image constructed strongly depends on the binarized image used as a constraint during the dilation process of the skeleton. The ground truth binarized images used for the DIBCO competition series are constructed with a modified procedure (Ntirogiannis et al., 2013) as illustrated in Figure 7. In this procedure, the conditional dilation step of the skeleton is constrained only by Canny edge image, without any initial binarized image.

Based on the preliminary experiments, it is expected to obtain a good initial binarized image as the input to the next process of ground truth creation (Kesiman et al., 2015a). The initial binarization method used in the stage of construction of skeletonized ground truth image should be able to generate an optimal and acceptable 'good enough' skeleton which detects and keeps the form of the characters. An image of skeleton generated in this step will facilitate the manual correction process. More the skeleton is correct, more the manual process is easier and faster. For a nondegraded palm leaf manuscript, a simple global thresholded binarization method is sufficient to generate an acceptable binarized image and optimal image of the skeleton. However, this method is not adapted to degraded palm leaf manuscripts. Figure 8 shows some examples of the skeletonized image generated

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 243 of 404

with Matlab standard function bwmorph[2] from different binarized images using different binarization methods. Influenced by the dried palm leaf texture, the stroke of characters in palm leaf manuscripts is thickened and widened. As a consequence, a lot of small short unuseful branches on the skeleton are generated. Because of the poor quality of the binarized and skeletonized image, the step of manual correction of the skeleton is very time consuming, it takes almost 8 hours for only one image of palm leaf manuscript. Therefore, in the case of degraded and low-quality palm leaf manuscript images, the study focused on the development of an initial binarization process for the construction of ground truth binarized images. One other important remark, superposing image of skeleton on the original image to guide the manual correction process is not enough. A priori knowledge of the form of ancient characters is mandatory to guarantee that the incomplete character skeleton can be completed in a more natural trace as the way how the characters have been originally written. The manual correction process should be done by a philologist or at least by a person who knows well how to write the ancient characters with a guide for the transcription of the manuscript provided by a philologist.

In order to overcome the binarization problem on degraded and low quality palm leaf manuscript image, the study of (Kesiman et al., 2015a) proposed a 'semi-local' concept. The idea of this method is to apply a powerful global binarization method on only precise local character area. The binarization scheme consists of several steps as illustrated in Figure 9. First, the edge detection with Prewitt operator is applied to get the initial surrounding area of the line-strokes of each character. Based on our visual observation, Prewitt leads to high edge response on the inner part of the characters, and it gives a good approximate area for the skeleton. Whereas Canny leads to high edge response on the outer side of text stroke, and it detects over sensitively the textural part of the palm leaf background. The grayscale image of the edge is then binarized with Otsu's method to get the first binarized image of the palm leaf manuscript. Median Filter is then applied to this binarized image in order to remove noise. After noise reduction, some characters might be affected and broken. A dilation process is applied to recover and reform the broken parts of the character. The method constructs the approximated character area using Run Length Smearing (RLS) method (Wahl et al., 1982). The smearing method should be done optimally, so the missing/undetected character area can be detected completely. The RLS in row wise will cover the missing area in horizontal strokes character line, meanwhile the RLS in column wise will cover the missing area in vertical strokes character line. The output of those steps is a binarized image with an approximated character area in black, and the background area in white. The next step is the main concept of this scheme. Otsu's binarization method is applied for the second time, but locally only within a limited character area, defined by each connected component from the first binarized image generated (Figure 10).

After the initial binarization process, the method finally performs a morphological-based thinning method to get the skeleton of the character. The thinned image normally still have the unwanted branch, so it applies a morphological-based pruning method to the thinned characters image. A pruning method for the skeleton is effective to remove spurious unwanted parts of the skeleton, and it makes the manual correction process of the skeleton faster. Figure 11 shows a sample of image sequence as the result of our specific scheme.

[2] http://fr.mathworks.com/help/images/ref/bwmorph.html.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 244 of 404

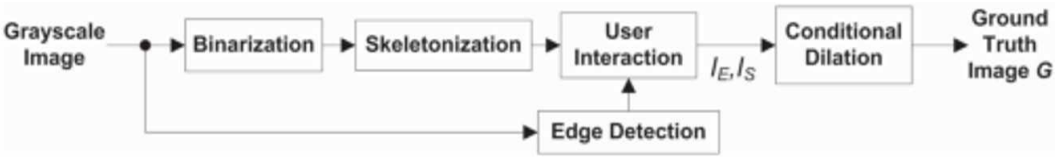


Figure 7. Ground truth construction procedure used for DIBCO series (Ntirogiannis et al., 2013).

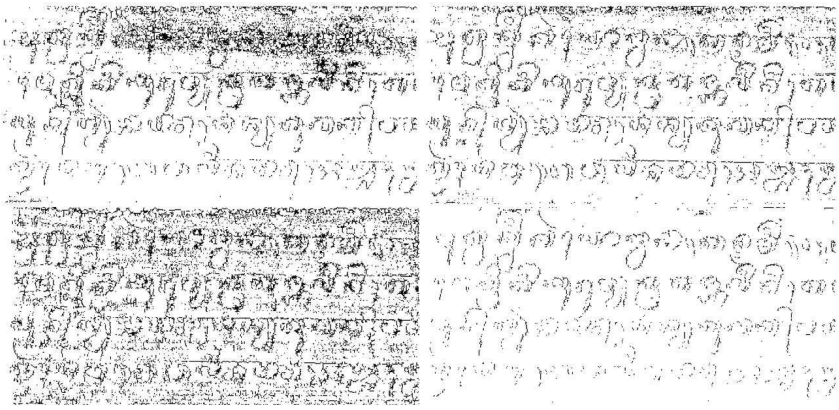


Figure 8. Examples of image of skeleton (left to right and up to bottom) (Kesiman et al., 2015a) generated from binarized image of Otsu (Pratikakis et al., 2013; Messaoud et al., 2011), Niblack (Khurshid et al., 2009; Rais et al., 2004; Gupta et al., 2007; He et al., 2005; Feng and Tan, 2004), Rais (Rais et al., 2004), and NICK (Khurshid et al., 2009).

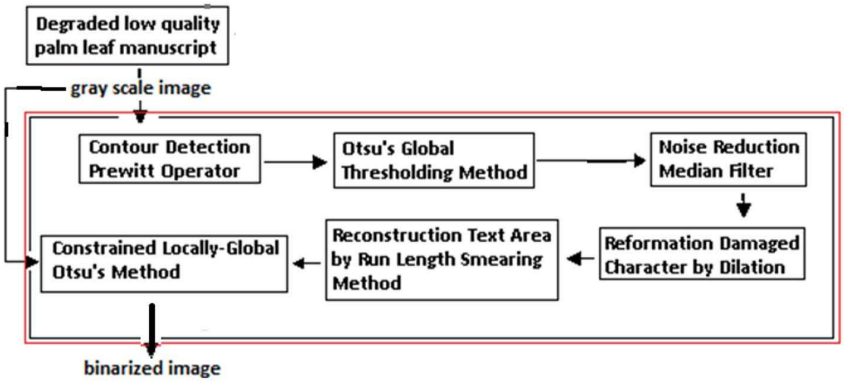


Figure 9. Semi-local binarization scheme (Kesiman et al., 2015a).

The goodness of the results can only be estimated qualitatively by examining the results. Based on visual criteria, the proposed scheme provides a good initial image of skeleton with respect to image quality and preservation of meaningful textual character information.

We experimentally tested the framework for the construction of ground truth binarized image for nondegraded and degraded low-quality palm leaf manuscript images (Kesiman et al., 2015a). For this initial experimental study, we only used the available sample scanned

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 245 of 404

images from Museum Bali, Museum Gedong Kertya, and from private family collection. The manuscripts were written on both sides, but there was no back-to-front interference observed.



Figure 10. Examples of extracted character area (on the left) and their semi-local binarization result (on the right) (Kesiman et al., 2015a).



Figure 11. Original sample image, and sequence sample image of Prewitt, Otsu, Median Filter, Dilation, RLS Row, RLS Col, Local Otsu, Thinning, Pruning, Superposed Skeleton on Original Image (Kesiman et al., 2015a).

For nondegraded palm leaf manuscripts, we used a simplest and the most conventional global thresholding method with a proper threshold selected manually to obtain the initial binarized image. With this initial binarized image, it is already sufficient to obtain an acceptable skeletonized image. We performed the manual correction of the skeleton, guided by the transcription of the manuscript provided by a philologist, to finally obtain the skeleton ground truth of the manuscript. Figure 12 shows a snapshot of a simple prototype with user friendly interface that we developed and used to facilitate the manual correction process. We finally constructed the ground truth image by dilating the corrected skeleton image, constrained by the Canny edge image and the initial binarized image from Otsu's global method. We use Otsu's global method instead of the same global fixed thresholding method used in our skeleton ground truth construction because we need a complete connected component of all characters detected on the binarized image. Other binarization methods can also be used, for example, Niblack's method or the multi resolution version of Otsu's method (Gupta et al., 2007). They also provide a satisfactory preliminary binarized image. Figure 13 shows an example of final ground truth image from a nondegraded palm leaf manuscript image. It is visually an acceptable estimated ground truth image for the

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 246 of 404

manuscript.

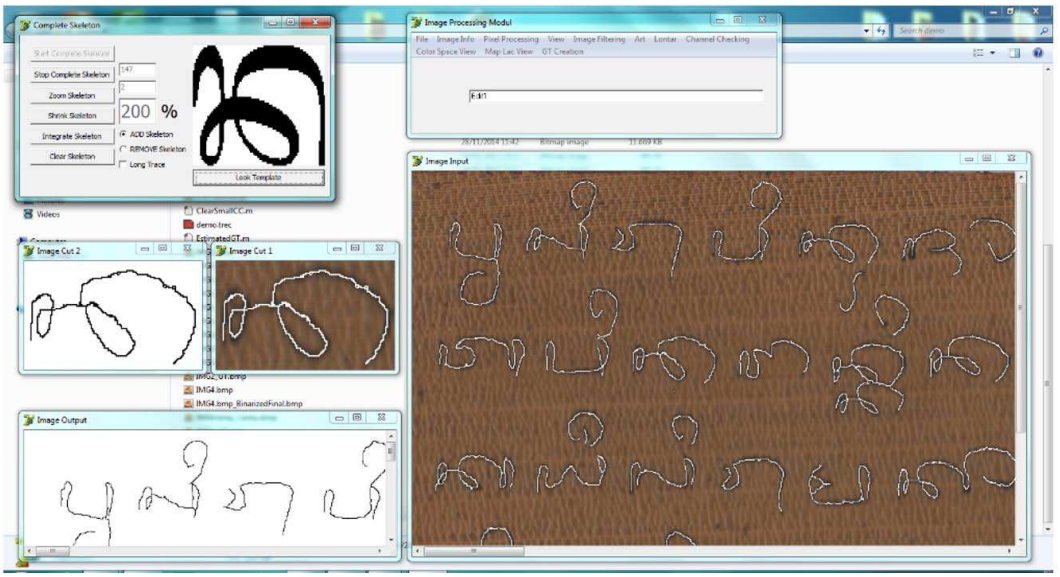


Figure 12. Snapshot of prototype interface used for manual correction of skeleton (Kesiman et al., 2015a).



Figure 13. Estimated ground truth of a nondegraded palm leaf manuscript image (Kesiman et al., 2015a).

For degraded low-quality palm leaf manuscript images, we applied our proposed specific binarization scheme by defining the optimal value of parameters based on our empirical experiments as follows: filter size 3x3 for Median Filter, square structuring element size 3x3 for Dilation, smearing 3 pixels in row and 3 pixels in column for RLS Method, and pruning the branch of 2 pixels. We performed the manual correction of the skeleton, guided by the transcription of the manuscript provided by a philologist to obtain the skeleton ground truth image of the manuscript. Figure 14 shows an example of a low-quality palm leaf manuscript and the skeleton ground truth image. We first experimented the construction of estimated ground truth image by applying a constraint of Canny edge image and an initial binarized image. For example, we used the binarized image from Niblack's method or the multi-resolution version of Otsu's method as the constraint. The estimated

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 247 of 404

ground truth image really depends on the initial binarized image used as a constraint. We then experimented the construction of ground truth image without any initial binarized image as a constraint. The result is shown in Figure 15. Based on visual criteria, the proposed algorithm seems to achieve a better-estimated ground truth image with respect to image quality and preservation of meaningful textual character information. Some other results of ground truth binarized image for degraded low-quality palm leaf manuscript images are shown in Figure 16.

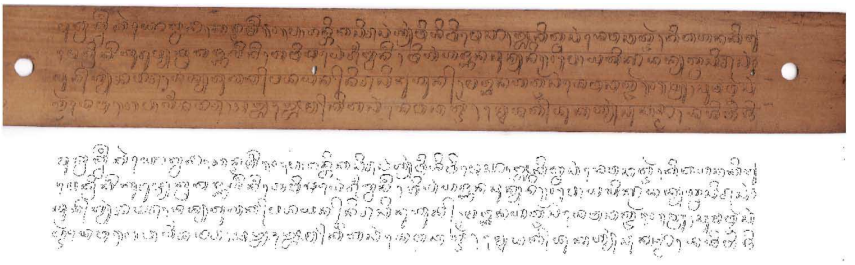


Figure 14. Original Image and the skeleton ground truth (Kesiman et al., 2015a).

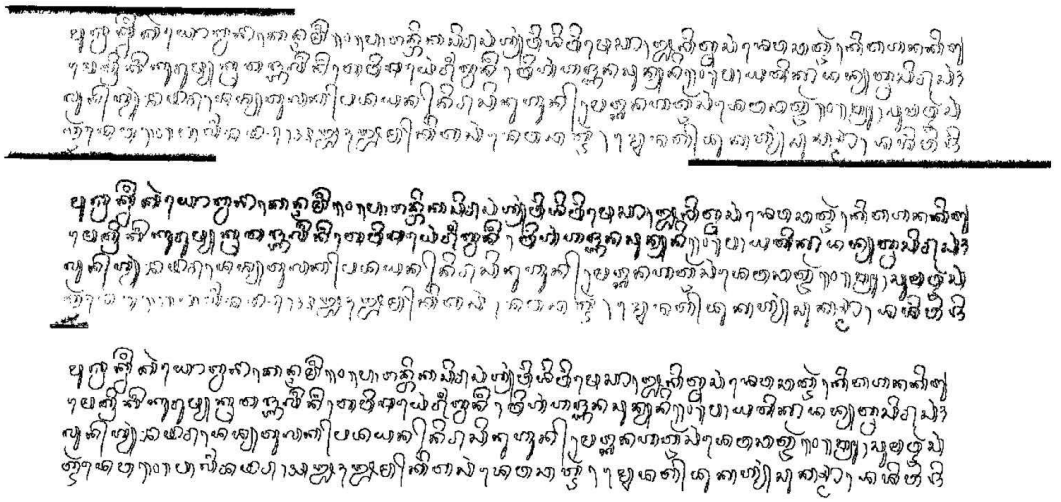


Figure 15. Ground truth image constructed with an initial binarized image of Niblack's method, Multi Resolution Otsu's method, and without any constraint of initial binarized image (Kesiman et al., 2015a).

2.3. Analysis of Ground Truth Binarized Image Variability

Regarding the human intervention in ground truthing process, the subjectivity effect on the construction of ground truth binarized image needs to be analyzed and reported. The work of (Smith, 2010) and (Smith and An, 2012) analyzed the binarization ground truthing and the effect of ground truth on image binarization of DIBCO binarized image dataset (Gatos et al., 2011). The study stated that the different choice of binarization ground truth affects the binarization algorithm design and the performance can vary significantly depending on

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 248 of 404

the choice of ground truth. In this section, we present an experiment in a real condition to analyze the human intervention subjectivity on the construction of ground truth binarized image and to measure quantitatively the ground truth variability of palm leaf manuscript images with different binarization evaluation metrics (Kesiman et al., 2015b)(Kesiman et al., 2015b). This experiment measures the difference between two ground truth binarized images from two different ground truthers.

The sample images used in this experiment are 47 images randomly selected from the palm leaf manuscript corpus of AMADI Project (see section "The palm leaf manuscript corpus and the digitization process"). In this experiment, we adopted a semi-automatic framework for the construction of ground truth binarized images which was described in section "The construction of ground truth binarized images". But, in order to measure the variability of human subjectivity in our ground truth creation, in this experiment, we did not apply any initial binarization and skeletonization methods. The skeletonization process is completely performed by human. The skeleton drawn manually by user is dilated until Canny edges intersect each binarized component of the dilated skeleton in a ratio of 0.1. This value of minimal ratio between number of pixels in intersection of Canny edge and number of pixels of the dilated skeleton is found based on our empirical experiment and observation on the thickness of the character stroke in our manuscripts.

As presented in (Smith, 2010), three metrics of binarization evaluation proposed in the DIBCO 2009 contest (Gatos et al., 2011) are used in this analysis to measure the difference between two ground truth binarized images from two different ground truthers. Those three metrics are F-Measure (FM), Peak SNR (PSNR), and Negative Rate Metric (NRM) (Kesiman et al., 2015b). The value of FM and PSNR when we assumed the image drawn by the first ground truther as ground truth image will be the same with the value of FM and PSNR when we assumed the vice versa, the image drawn by the second ground truther as ground truth image. The value of NRM when we assumed the image drawn by the first ground truther as ground truth image will not be the same with the value of NRM when we assumed the image drawn by the second ground truther as ground truth image. In this case, we calculated two value of NRM: NRM1 and NRM2. A higher F-measure and PSNR indicates a better match. A lower NRM indicates a better match.

For this experiment, 70 students were asked to trace manually the skeleton of the Balinese character found in palm leaf manuscript image with PixLabeler tool (Saund et al., 2009). One student worked with two different images, and one image was ground truthed by two different students. These two manually skeletonized image will be re-skeletonized with Matlab function bwmorph[3] to make sure that the skeleton is one pixel wide for the next process of automatic ground truth estimation with conditional dilation and Canny edge constraint. Figure 17 shows the scheme diagram of our experiment. Figure 18 shows some samples images as the result example of this experiment.

By observing visually the two skeletonized image created by two different ground truthers, we can see how different are the results of the two ground truthers in choosing the trace of the character skeleton. All the broken parts of in image of intersection between two skeletonized images show the different skeleton traces between two ground truthers. And all the double-lined parts in the image of union between two skeletonized images show how

[3]http://fr.mathworks.com/help/images/ref/bwmorph.html.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 249 of 404

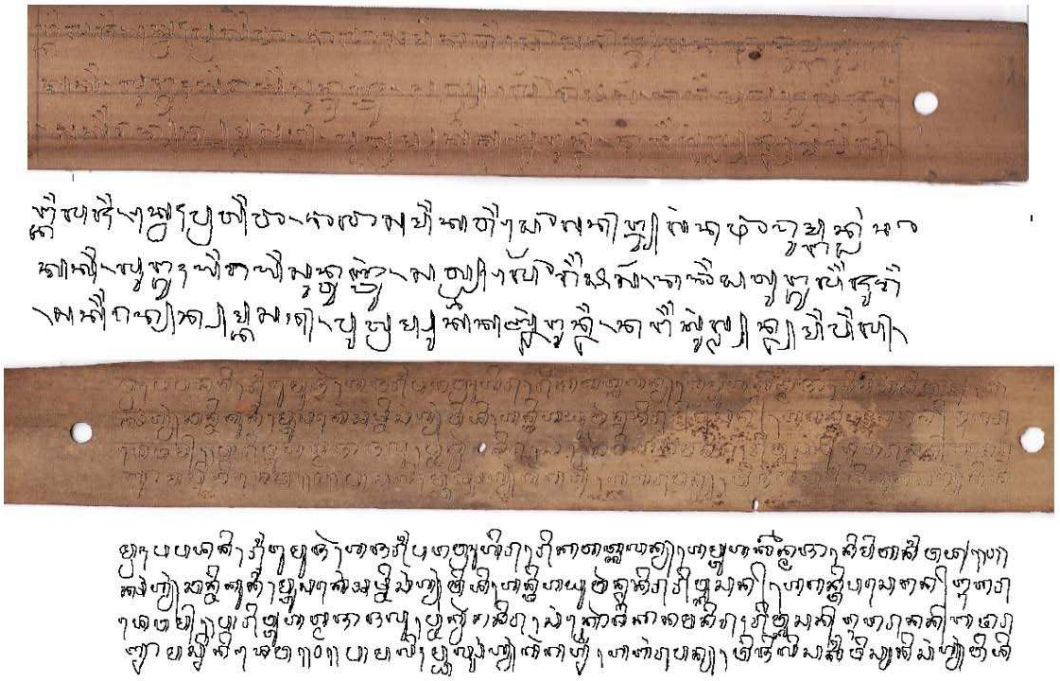


Figure 16. Two palm leaf manuscript images with their ground truth binarized images (Kesiman et al., 2015a).

far the different are the positions of the skeleton traced by two ground truthers.

First, we measured the variability between two skeletonized ground truthed images manually drawn by two different ground truthers (Table 1) (Kesiman et al., 2015b). The wide range between the maximum and the minimum value and also the mean and variance value of all three binarization evaluation metrics from 47 images show that there is a large variability between the ground truthers for each image.

Table 1. Variability between two manually skeletonized ground truthed image (Kesiman et al., 2015b)

Comparison metric	FM	NRM1	NRM2	PSNR
Maximum	58,945	0,371	0,458	60,166
Minimum	14,058	0,209	0,209	26,882
Mean	41,459	0,302	0,303	33,196
Variance	77,764	0,002	0,003	60,083

We then measured the variability between the two ground truth binarized images automatically estimated from two different manually skeletonized images for each image of the manuscript. Table 2 illustrates this variability (Kesiman et al., 2015b). The wide range between the maximum and the minimum value and also the mean and variance value of all three binarization evaluation metrics show that there is still a large variability between the estimated ground truth images for each image.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 250 of 404

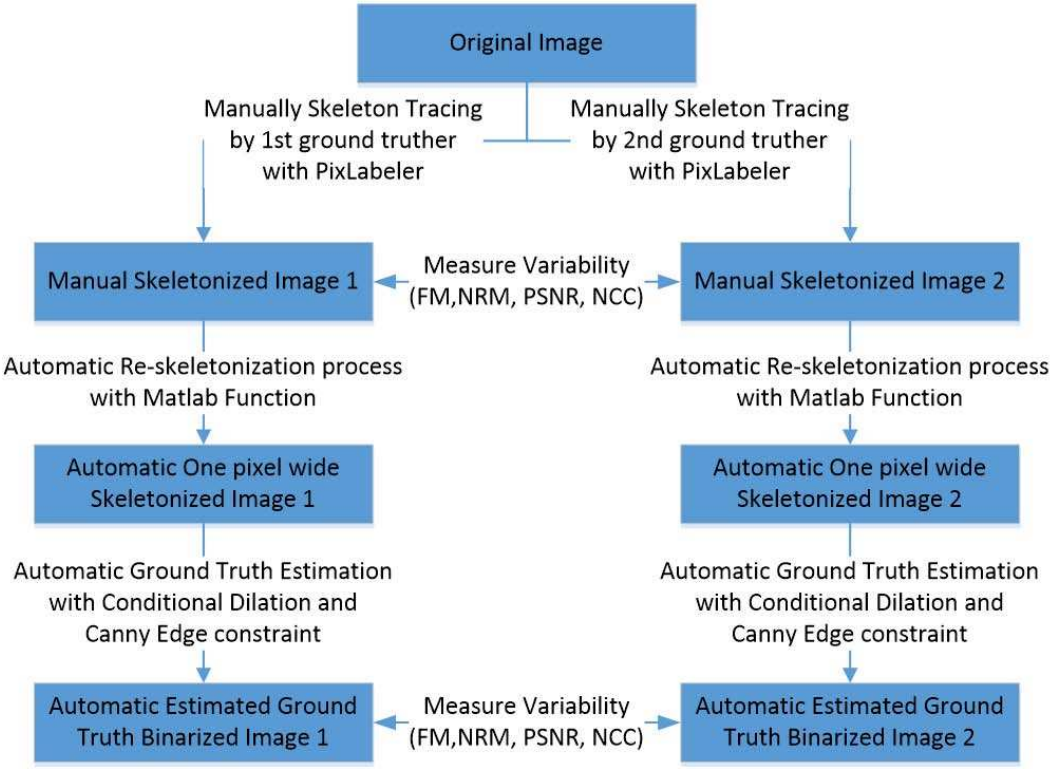


Figure 17. Scheme diagram of experiment (Kesiman et al., 2015b).

Table 2. Variability between two ground truthed image automatically estimated from two different manually skeletonized image (Kesiman et al., 2015b)

Comparison metric	FM	NRM1	NRM2	PSNR
Maximum	74,731	0,309	0,446	59,196
Minimum	18,615	0,128	0,13	23,961
Mean	59,556	0,214	0,215	31,11
Variance	89,88	0,002	0,003	61,383

By comparing the value of binarization evaluation metrics between the two manually skeletonized ground truth images (Table 1) and between the two automatic estimated ground truth images (Table 2), we can see that the variability of two ground truth images in F-hMeasure and NRM for all images decreases after the estimation ground truth process. The value of PSNR decreases because the number of different foreground-background pixels between the two estimated ground truth images also increases after the automatic estimation process, not only the number of common foreground pixels from the two estimated ground truth images. Figure 19 to Figure 22 shows that the ground truth estimation process tends to decrease the variability between two ground truthers to produce a better match between two ground truth images.

We also tested and estimated the ground truth binarized image from the union of two

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 251 of 404

skeleton images manually drawn by two differents ground truthers (see exemple in Figure 18(e)). The variability between this estimated union ground truth image with two other estimated ground truth images from each ground truther is then measured. Table 3 and Table 4 illustrate the results of the comparison metric for all images in experiment (Kesiman et al., 2015b). The ground truth image, estimated from the union of two skeleton images, indicates a better match with two other ground truth images from two different ground truthers.

Table 3. Variability between ground truth image estimated from union of two skeleton images with ground truth image estimated from the first ground truther (Kesiman et al., 2015b)

Comparison metric	FM	NRM1	NRM2	PSNR
Maximum	89,758	0,076	0,418	67,095
Minimum	27,823	0,038	0,064	29,854
Mean	80,539	0,066	0,132	37,759
Variance	71,677	0	0,003	70,775

Table 4. Variability between ground truth image estimated from union of two skeleton images with ground truth image estimated from the second ground truther (Kesiman et al., 2015b)

Comparison metric	FM	NRM1	NRM2	PSNR
Maximum	94,182	0,09	0,227	65,188
Minimum	66,806	0,025	0,035	30,815
Mean	81,155	0,067	0,129	37,816
Variance	17,054	0	0,001	63,464

Based on our data survey after the experiment with all ground truthers, we have observed and remarked some facts on the ground truth creation of palm leaf manuscripts as follows: The Balinese alphabet found in the manuscripts are not daily used by the ground truthers. Most of the ground truthers learned those alphabets in their elementary school until their junior or senior high school, but they never re-used those alphabets after the classroom learning process. There are some characters of the alphabet that they have never seen before. For those kinds of characters, the ground truthers could not make a smooth and natural trace of the character skeleton. Regarding the variability of ground truth images produced in this experiment, we suggest that this kind of important fact or condition should be always taking into account in every ground truthing process of ancient manuscript project. The time needed to semi-manually corrected the skeleton of the image from an initial automatic method can be much greater than making the skeleton totally manual started from zero. In our first trial experiment, we need 4 until 6 hours to corrected the semi-automatic generated skeleton. It is due to the physical characteristics of the manuscripts which make the binarizing and skeletonizing method do not tend to produce the optimal good skeleton of the characters. We finally decided to make it totally manual, and it takes between 2 to 3 hours to trace the skeleton started from zero.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 252 of 404

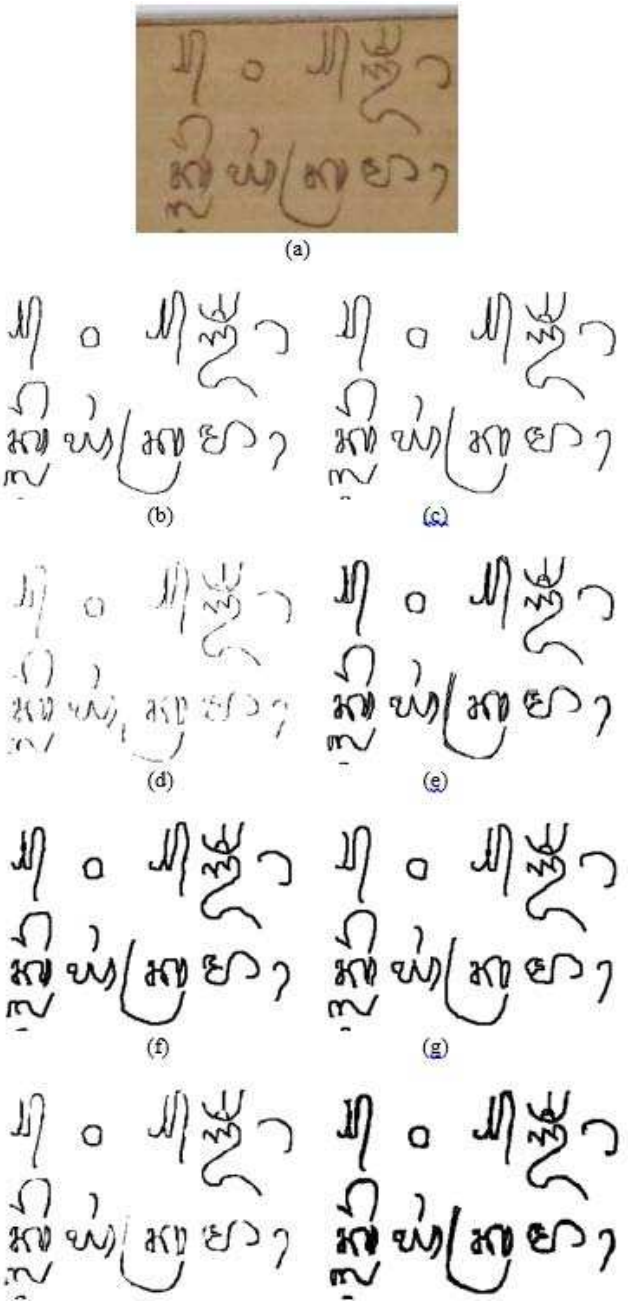


Figure 18. Example of ground truth binarized image from the experiment: (a) original image, (b) skeletonized image by 1st ground truther, (c) skeletonized image by 2nd ground truther, (d) image intersection between (b) and (c), (e) image union between (b) and (c), (f) estimated ground truth binarized image from (b), (g) estimated ground truth binarized image from (c), (h) image intersection between (f) and (g), (i) image union between (f) and (g) (Kesiman et al., 2015b).

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 253 of 404

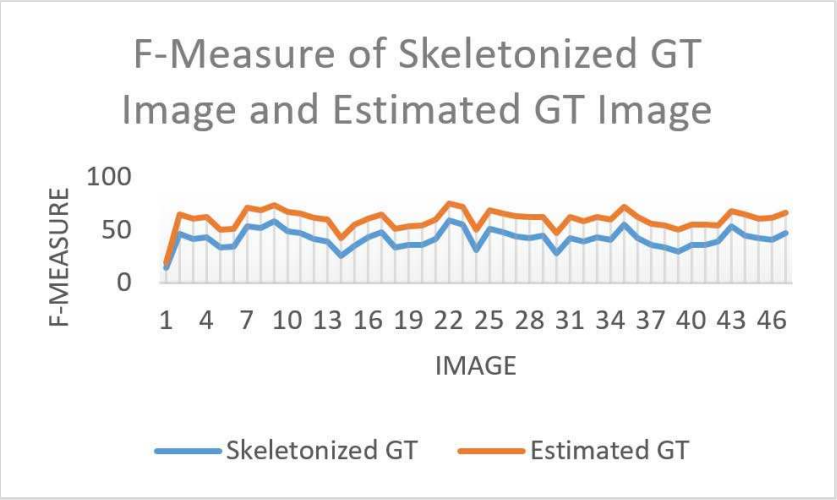


Figure 19. Comparison of F-Measure between two skeletonized ground truth image and between two estimated ground truth images (Kesiman et al., 2015b).

From the result of this experiment, we proved that the human subjectivity has a great effect in producing a great variability of ground truth binarized image. This phenomenon becomes much more visible when we are working on the binarization process of ancient type document or manuscript where the physical characteristics and conditions of the manuscript are not good enough or it is still hard to be ground truthed even by human. The method of binarization evaluation by comparing and measuring pixel-by-pixel with a ground truth binarized image should be re-evaluated to avoid the great bias from human subjectivity. Some other measures should be proposed to evaluate the binarization process of document image of ancient manuscripts.

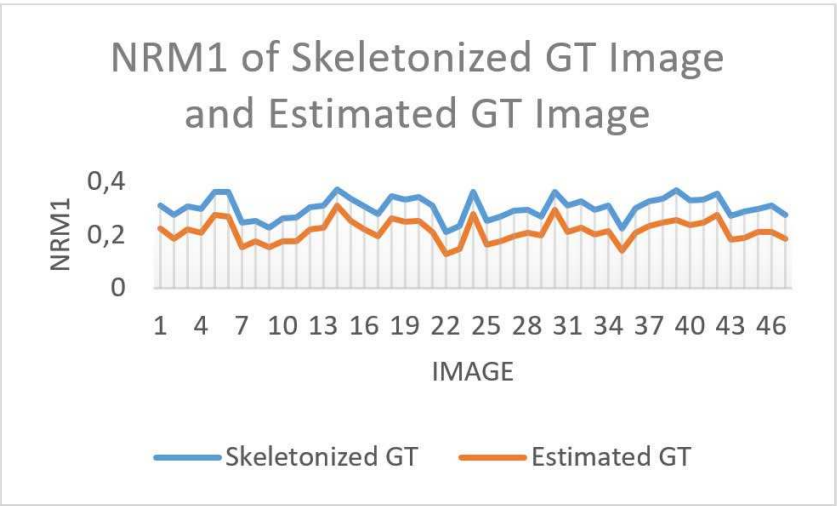


Figure 20. Comparison of NRM1 between two skeletonized ground truth image and between two estimated ground truth images (Kesiman et al., 2015b).

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 254 of 404

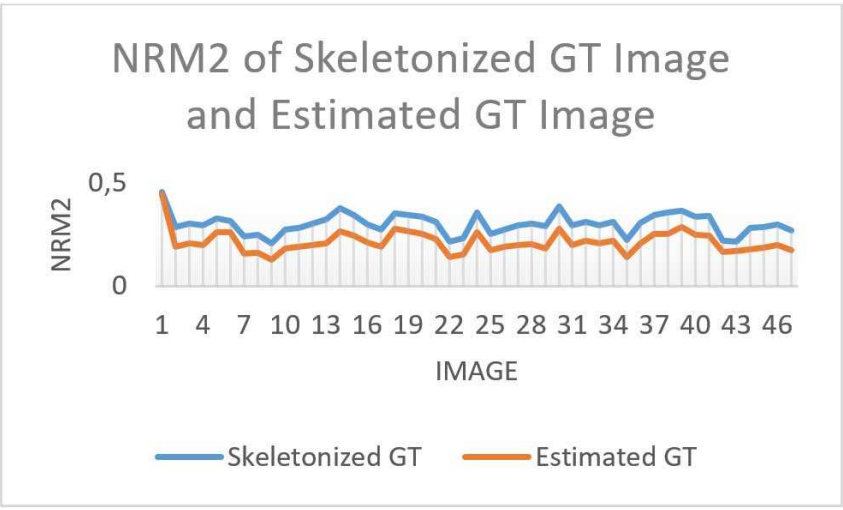


Figure 21. Comparison of NRM2 between two skeletonized ground truth image and between two estimated ground truth images (Kesiman et al., 2015b).

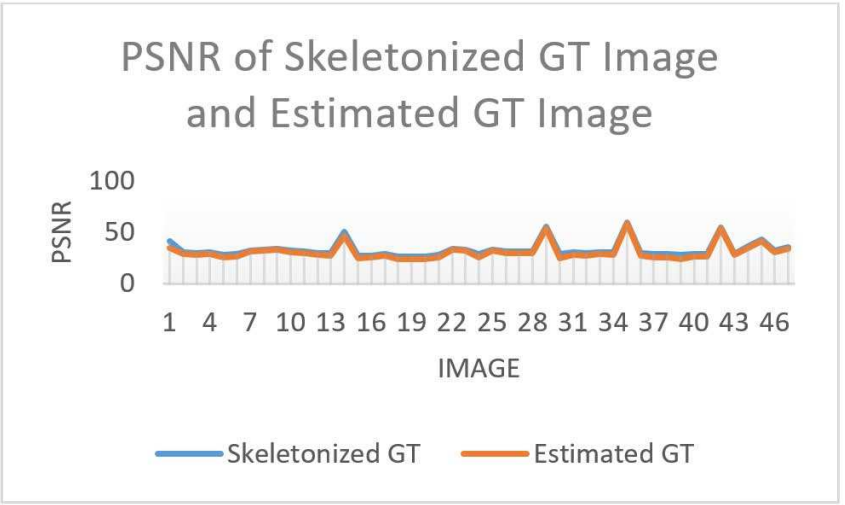


Figure 22. Comparison of PSNR between two skeletonized ground truth image and between two estimated ground truth images (Kesiman et al., 2015b).

3. Isolated Character Recognition

Isolated handwritten character recognition (IHCR) has been the subject of intensive research during the last three decades. Some IHCR methods have reached a satisfactory performance, especially for Latin script. However, development of IHCR methods for other various new scripts remains a major task for researchers. For example, the IHCR task for historical documents discovered in the palm leaf manuscripts.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 255 of 404

IHCR system is one of the most demanding systems which has to be developed for the collection of palm leaf manuscript images. Using an IHCR system will help to transcript these ancient documents and translate them to the current language. Usually, an IHCR system consists of two main steps: feature extraction and classification. The performance of an IHCR system greatly depends on the feature extraction step. The goal of feature extraction is to extract information from raw data which is most suitable for classification purpose (Aggarwal et al., 2015). Many feature extraction methods have been proposed to perform the character recognition task (Arica and Yarman-Vural, 2002; Blumenstein et al., 2003; Aggarwal et al., 2015; Kumar, 2010; Bokser, 1992; Hossain et al., 2012; Fujisawa et al., 1999; Jin et al., 2009; Rani and Meena, 2011). These methods have been successfully implemented and evaluated for recognition of Latin, Chinese and Japanese characters as well as digit recognition. However, only a few systems are available in the literature for other Asian scripts recognition. For example, some of the works are for Devanagari script (Kumar, 2010; Ramteke, 2010), Gurmukhi script (Aggarwal et al., 2015; Lehal and Singh, 2000; Sharma and Jhajj, 2010; Siddharth et al., 2011), Bangla script (Hossain et al., 2012), and Malayalam script (Ashlin Deepa and Rajeswara Rao, 2014). Those documents with different scripts and languages surely provide some new research problem, not only because of the different shapes of characters but also because the writing style for each script differs: the shape of the characters, character positions, separation or connection between the characters in a text line.

Each feature extraction method has its own advantages or disadvantages over other methods. In addition, each method may be specifically designed for some specific problem. Most of feature extraction methods, extract the information from binary image or grayscale image (Kumar, 2010). Some surveys and reviews on features extraction methods for character recognition were already reported (Trier et al., 1996; Kumar, 2011; Neha J. Pithadia, 2015; Pal et al., 2012; Pal and Chaudhuri, 2004; Govindan and Shivaprasad, 1990). Choosing efficient and robust feature extraction methods plays a very important role to achieve high recognition performance in an IHCR and OCR (Aggarwal et al., 2015). The performance of the system depends on a proper feature extraction and a correct classifier selection (Hossain et al., 2012). It is experimentally reported that to improve the performance of an IHCR system, the combination of multi features is recommended (Trier et al., 1996). Our objective is to find the combination of feature extraction methods to recognize the isolated characters of Balinese scripts on palm leaf manuscript images.

In this work, first, we investigated and evaluated some most commonly used features for character recognition: histogram projection (Kumar, 2010; Hossain et al., 2012; Ashlin Deepa and Rajeswara Rao, 2014), celled projection (Hossain et al., 2012), distance profile (Bokser, 1992; Ashlin Deepa and Rajeswara Rao, 2014), crossing (Kumar, 2010; Hossain et al., 2012), zoning (Blumenstein et al., 2003; Kumar, 2010; Bokser, 1992; Ashlin Deepa and Rajeswara Rao, 2014), moments (Ramteke, 2010; Ashlin Deepa and Rajeswara Rao, 2014), some directional gradient based features (Aggarwal et al., 2015; Fujisawa et al., 1999), Kirsch Directional Edges (Kumar, 2010), and Neighborhood Pixels Weights (NPW) (Kumar, 2010). Secondly, based on our preliminary experiment results, we proposed and evaluated the combination of NPW features applied on Kirsch Directional Edges images, with Histogram of Gradient (HoG) features and zoning method. Two classifiers: k-NN (k-Nearest Neighbor) and SVM (Support Vector Machine) are used in our

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 256 of 404

experiments. This section will only briefly describe the feature extraction methods which were used in our proposed combination of features. For more detail description of other commonly used feature extraction methods which were also evaluated in this experimental study, please refer to references mentioned above.

3.1. Kirsch Directional Edges

Kirsch edges method is a non-linear edge enhancement (Kumar, 2010). Let $Ai(i = 0, 1, 2, ..., 7)$ be the eight neighbors of the pixel (x,y), i is taken as modulo 8, starting from top left pixel at the moving clock-wise direction. Four directional edge images are generated (Figure 23) by computing the edge strength at pixel (x,y) in four (horizontal, vertical, left diagonal, right diagonal) directions, defined as G_H, G_V, G_L, G_R, respectively (Kumar, 2010). They can be denoted as bellow:

$$G_H(x, y) = max(|5S_0 - 3T_0|, |5S_4 - 3T_4|) \quad (1)$$
$$G_V(x, y) = max(|5S_2 - 3T_2|, |5S_6 - 3T_6|) \quad (2)$$
$$G_R(x, y) = max(|5S_1 - 3T_1|, |5S_5 - 3T_5|) \quad (3)$$
$$G_L(x, y) = max(|5S_3 - 3T_3|, |5S_7 - 3T_7|) \quad (4)$$

where S_i and T_i can be computed by:

$$S_i = A_i + A_{i+1} + A_{i+2} \quad (5)$$

$$T_i = A_{i+3} + A_{i+4} + A_{i+5} + A_{i+6} + A_{i+7} \quad (6)$$

Each directional edge image is thresholded to produce a binary edge image. The binary edge image is then partitioned into N smaller regions. Then, edge pixel frequency in each region is computed to produce the feature vector. In our experiments, we computed Kirsch feature from grayscale image with 25 smaller regions to produce a 100 dimensions feature vector. Based on the empirical tests for our dataset, the Kirsch edge image can be optimally thresholded with a threshold value of 128. The feature value is then normalized by the maximum value of edge pixel frequency from all regions.

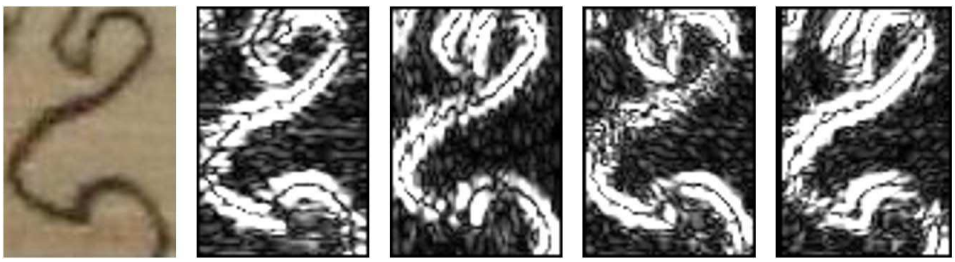


Figure 23. Four directional Kirsch edge images [6].

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 257 of 404

3.2. Neighborhood Pixels Weights

Neighborhood Pixels Weight (NPW) was proposed by Satish Kumar (Kumar, 2010). This feature may work on binary as well as on gray images. NPW considers four corners of neighborhood for each pixel: top left, top right, bottom left, and bottom right corner. The number of neighbors considered on each corner is defined by the value of layer level (see Figure 24). Level 1 considers only pixels in layer 1 on each corner (1 pixel), level 2 considers pixels in layer 1 and 2 (4 pixels), and level 3 considers pixels in all layers (9 pixels). In the case of the binary image, the weight value on each corner is obtained by counting the number of pixel character, divided by a total number of neighborhood pixels on that corner. For the grayscale image, the weight value on each corner is obtained by summing the gray level of all neighborhood pixels, divided by maximum possible weight due to all neighborhood pixels on that corner (nb_neighborhood_pixels x 255). Four weighted plans are constructed for each corner from the weighted value of all pixels of the image. Each plane is divided into N smaller regions, and the average weight of each region is computed. The feature vector is finally constructed from the average weight of each region from each plane (N x 4 vector dimension).

3	3	3		3	3	3
3	2	2		2	2	3
3	2	1		1	2	3
			P			
3	2	1		1	2	3
3	2	2		2	2	3
3	3	3		3	3	3

Figure 24. Neighborhood pixels for NPW features (Kesiman et al., 2016c).

In our experiments, we computed NPW features in level 3 neighborhood with 25 smaller regions (N=25) to produce 100 dimensions feature vector. The feature value is normalized by the maximum value of average weight from all regions. We tested the performance of NPW feature for both binary and grayscale image.

3.3. Histogram of Gradient

The gradient is a vector quantity comprising of magnitude as well as directional component computed by applying its derivatives in both horizontal and vertical directions (Aggarwal et al., 2015). The gradient of an image can be computed either by using, for example, Sobel, Roberts or Prewitt operator. The gradient strength and direction can be computed from the gradient vector. Gradient feature vector used in (Aggarwal et al., 2015) is formed by accumulating the gradient strength separately along different directions.

To compute the histogramme of gradients (HoG), first, we calculate the gradient magnitude and gradient direction of each pixel of the input image. The gradient image is then divided into some smaller cells, and in each cell, we generate the histogram of directed

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 258 of 404

gradient by assigning the gradient direction of each pixel into certain range of orientation bin which are evenly spread over 0 to 180 degrees or 0 to 360 degrees (Figure 25 & 26). The histogram cells are then normalized with a larger overlap-connected blocks.

The final HoG descriptor is then generated from all concatenated vector of the histogram after the block normalization process. For our experiments, we used the HoG implementation of VLFeat[4]. We computed HoG feature from grayscale image with the cell size of 6 pixels and with 9 different orientations to produce a 1984 dimensions feature vector.

3.4. Zoning

Zoning is computed by dividing the image into N smaller zones: vertical, horizontal, square, diagonal left and right, radial or circular zone (see Figure 27). The local properties of image are extracted on each zone. Zoning can be implemented for binary image and grayscale image (Kumar, 2010). For example, in binary image, the percentage density of character pixels in each zone is computed as local feature (Bokser, 1992). In grayscale image, the average of gray value in each zone is considered as local feature (Ashlin Deepa and Rajeswara Rao, 2014). Zoning can be easily combined with other feature extraction methods (Hossain et al., 2012), for example in (Blumenstein et al., 2003). In our experiments, we computed zoning with 7 zone types (zone width or zone size = 5 pixels) and combined them into a 205 dimensions feature vector. We also tested the zoning feature on the image of the skeleton.

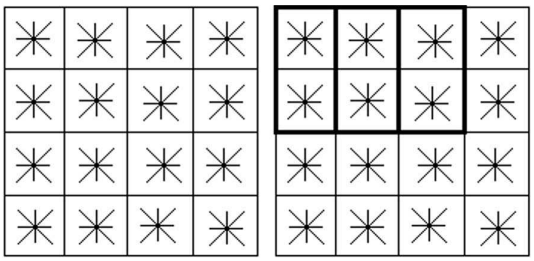


Figure 25. An image with 4x4 oriented histogram cells and 2x2 descriptor blocks overlapped on 2x1 cells (Kesiman et al., 2016c).



Figure 26. The representation of the array of cells HoG (Kesiman et al., 2016c).

[4]http://www.vlfeat.org/api/hog.html.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 259 of 404

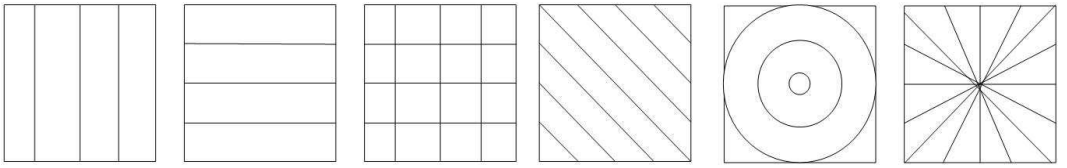


Figure 27. Type of Zoning (from left to right: vertical, horizontal, block, diagonal, circular, and radial zoning) (Kesiman et al., 2016c).

3.5. Our Proposed Combination of Features

After evaluating the performance of 10 individual feature extraction methods, we found that the HoG features, NPW features, Kirsch features, and Zoning method provide a good enough result (see Table 10). We obtained 62,45% of recognition rate only by using Kirsch features. It means that the four directional Kirsch edge images already serve as a good feature discriminants for our dataset. The shape of Balinese characters is naturally composed of some curves. We can notice that Kirsch edge image is able to give the initial directional curve features for each character. On the other hand, NPW features have an advantage that it can be applied directly to gray level images. Our hypothesis is the four directional Kirsch edge images will provide a better feature discriminants for NPW features. Based on this hypothesis, we proposed a new feature extraction method by applying NPW on kirsch edge images. We call this new method as NPW-Kirsch (see Figure 28). Finally, we concatenate NPW-Kirsch with two other features, HoG and Zoning methods.

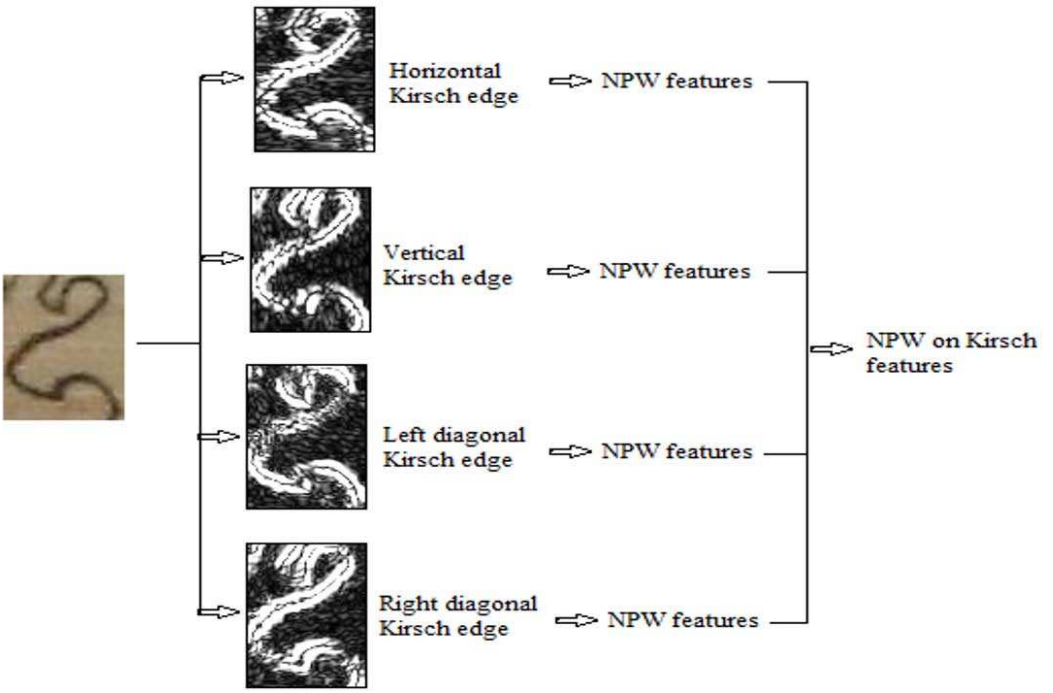


Figure 28. Scheme of NPW on Kirsch features (Kesiman et al., 2016c).

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 260 of 404

4. Word Spotting

Many works on word spotting methods have been reported for the last decade (Lee et al., 2012a; Dovgalecs et al., 2013; Rusinol et al., 2011; Rothacker et al., 2013a; Khayyat et al., 2013; Fischer et al., 2012; Rothacker et al., 2013b). The segmentation free word spotting method tries to spot the query word patch image given by the user by applying a sliding window on the document image (Dovgalecs et al., 2013; Rusinol et al., 2011; Rothacker et al., 2013a,b; Lee et al., 2012a). For each zone, the system measures the similarity between the query image based on some image features or descriptors. The training based word spotting method integrates the learning system to recognize the query word patch image on the document image (Rothacker et al., 2013a; Khayyat et al., 2013; Rothacker et al., 2013b). The system should be sufficiently trained by a collection of training data to achieve a better performance.

As the benchmark, most of the proposed word spotting methods were tested and evaluated on the collection of document images which were printed or handwritten on the paper with Latin script in English (Lee et al., 2012a; Dovgalecs et al., 2013; Rusinol et al., 2011; Rothacker et al., 2013b), for example, the well-known and widely used collection of George Washington document dataset[5]. Some methods were already proposed and evaluated to spot the word on the collection of document images with non-Latin scripts, for example, on Korean, Persian, Arabic documents and also on Indic scripts and languages (Rusinol et al., 2011; Rothacker et al., 2013a; Khayyat et al., 2013; Lee et al., 2012a). The writing style for each script differs in how to write and to join or separate a word in a text line.

Based on some surveys, an image feature which has been widely used to proceed the matching task on image retrieval and indexing systems is the Scale Invariant Features Transform (SIFT) (Dovgalecs et al., 2013; Rusinol et al., 2011; Lee et al., 2012a; Auclair et al., 2007; Almeida et al., 2009; Ledwich and Williams, 2004; Lowe, 2004). Based on the work of Rusiñol et. al. (Rusinol et al., 2011) and Dovgalecs et. al. (Dovgalecs et al., 2013), we experiment a segmentation free and training free word spotting method for our multi-writer palm leaf manuscript images using Bag of Visual Words (BoVW). Our preliminary hypothesis is the powerful framework of BoVW, combined with Latent Semantic Indexing (LSI) (Rusinol et al., 2011; Deerwester et al., 1990), Longest Common Subsequence (LCS) (cor, 2001), and Longest Weighted Profile (LWP) (Dovgalecs et al., 2013). The segmentation free and training free word spotting method is more suitable to be used for palm leaf manuscript images, because as we already stated, words in Balinese script were not written separately, so in this case, word segmentation is not a trivial process for this collection.

4.1. Offline Feature Extraction of Manuscript Images with Bag of Visual Word (BoVW)

For each page of manuscript, we applied this procedure. 1) Keypoints detection with densely SIFT descriptors[6]: We densely calculated the SIFT descriptors every 5 pixels using squared region of 48 pixels. We experimentally found that this spatial parameter can optimally cover each character on the manuscript with the descriptor points (Figure 29).

[5] http://www.iam.unibe.ch/fki/databases/iam-historical-document-database/washington-database.
[6] http://www.vlfeat.org/api/sift.html.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 261 of 404

Each descriptor contains 128 feature values. 2) Descriptors removal based on gradient norm: We only kept 75% of descriptors with the highest gradient norm, and remove most of the descriptors in the manuscript background (Figure 30). 3) Descriptors quantization into codebook with K-Means Clustering[7]: We quantized all descriptors into 1500 clusters. 4) Visual words construction with codebook cluster: We assigned a cluster label for each keypoints (Figure 31). 5) Bag of Visual Words (BoVW) construction with Spatial Pyramid Matching (SPM) (Lazebnik et al., 2006) of visual word patches: We generated the histogram of visual words by sliding a patch of size 300 x 75 pixels, sampled at each 25 pixels (Figure 32 and Figure 33). This patch size covers sufficiently the average size of a word. Based on SPM level 2, the histogram was constructed for each patch from 3 spatial positions: total area of patch, left area of patch, and right area of patch (Figure 34). Those three histograms from 3 spatial positions of patch, each with 1500 bin values of 1500 label clusters, are then concatenated into one histogram feature with 4500 bin values (Figure 35) (as jth feature patch of ith page of manuscript (P_i^j)).

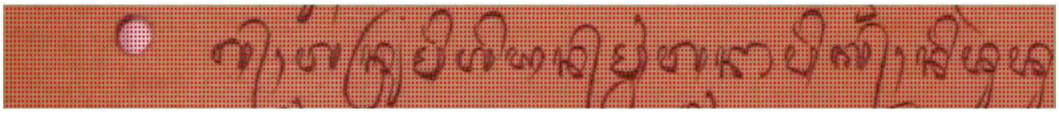


Figure 29. Densely detected SIFT descriptors.



Figure 30. SIFT descriptors with high gradient norm.



Figure 31. Visual words with codebook clusters.

4.2. Latent Semantic Indexing

To be able to retrieve relevant patches which do not contain a whole features of query, the use of Latent Semantic Indexing (LSI) was proposed (Rusinol et al., 2011). This semantic structure is defined by assigning to each patch descriptor a set of topics. All histogram features of visual word from one page of manuscript were formed into a matrices (A^i) of featurebypatch. This matrix then weighted by applying *tfidf* model. Singular Value Decomposition (SVD)[8] (Rusinol et al., 2011; Deerwester et al., 1990) was then applied to these matrices to reduce the feature space into a K-dimensional space. Matrice (A^i) was

[7]http://www.vlfeat.org/overview/kmeans.html.
[8]http://fr.mathworks.com/help/matlab/ref/svd.html

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 262 of 404

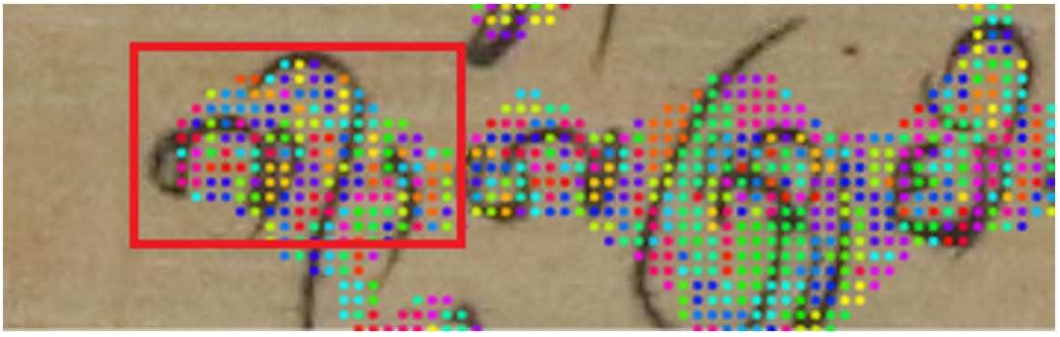


Figure 32. A patch of visual words

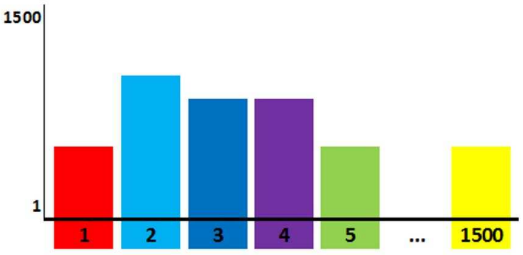


Figure 33. Histogram of a patch of visual words.

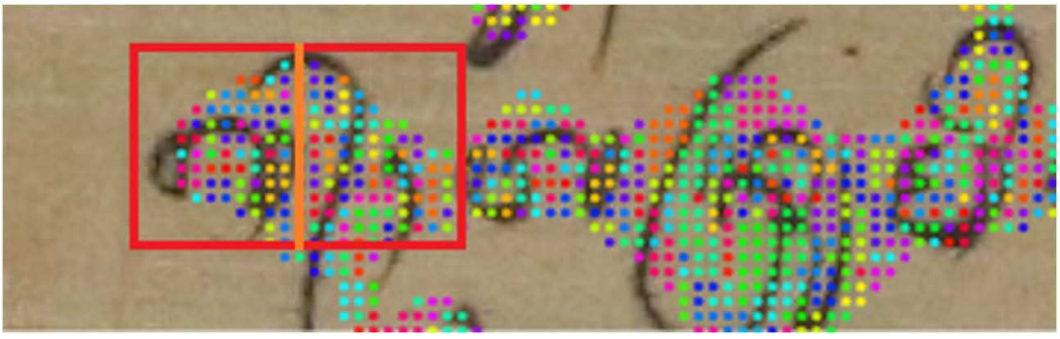


Figure 34. Spatial Pyramid Matching level 2 of a patch of visual words.

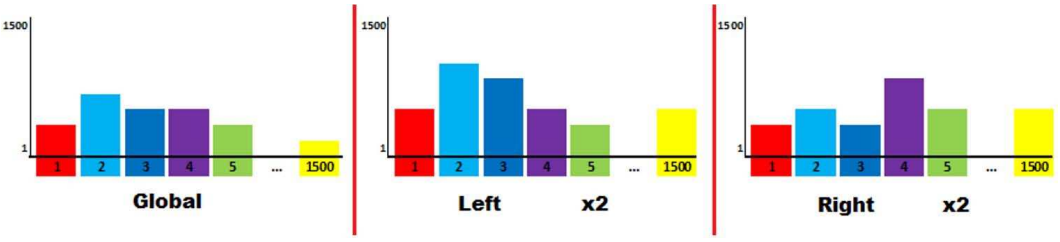


Figure 35. Histogram feature of a patch of visual words with SPM level 2.

decomposed into three matrices: U, S and V.

$$A^i; \hat{A}^i = U_K^i S_K^i (V_K^i)^T \quad (7)$$

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 263 of 404

where for the i[th] page of manuscript, $U_k \in R^{MxK}$, $S_k \in R^{KxK}$ and $V_k \in R^{NxK}$, M is the size of feature space, N is the number of patches in this i[th] page of manuscript. In our experiments, M=4500 and K=200. Too small K value will cause loss of important information. Each feature histogram of 4500 bin (P_j^i) is then transformed into a feature vector of 200 values ($\hat{P_j^i}$) by projecting the features into a topic space based on matrices U and S.

$$\hat{P_j^i} = (P_j^i)^T U_K^i (S_K^i)^{-1} \tag{8}$$

4.3. Online Feature Extraction of Query Image

For each query image, we applied exactly the same steps than the previous offline feature extraction process for each page of the manuscript, from the keypoints detection with densely SIFT descriptors until the descriptors quantization into a codebook with K-Means Clustering. But for this clustering process, we quantize all descriptors based on the clusters already defined for a page of manuscripts (Figure 36- 40). And for the SVD process, the histogram feature of 4500 bins of query image (Q^i) is then transformed into a feature vector of 200 values ($\hat{Q^i}$) by projecting the features into a K-dimensional space based on matrices U and S which are already generated for each page of manuscript.

$$\hat{Q^i} = (Q^i)^T U_K^i (S_K^i)^{-1} \tag{9}$$



Figure 36. Densely detected SIFT descriptors of a query image.



Figure 37. SIFT descriptors with high gradient norm of a query image.

4.4. Online Matching Between Query and Patches of Word on the Page of Manuscripts

For the matching process, the following method is applied. 1) Similarity measure with Cosine Distance: For each query image and the i[th] page of the manuscript, we measure the

Cerence Operating Company, IPR 2024
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 264 of 404

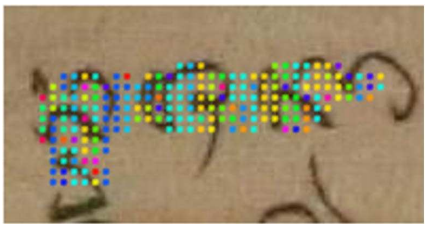


Figure 38. Visual words with codebook clusters of a query image.

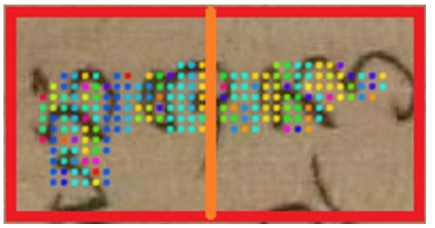


Figure 39. Spatial Pyramid Matching level 2 of a patch of visual words of a query image.

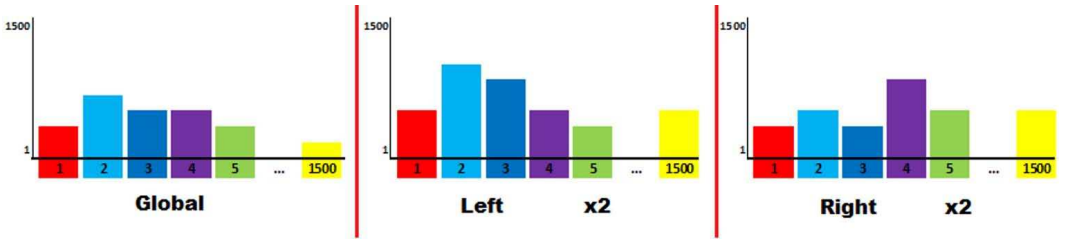


Figure 40. Histogram feature of a patch of visual words with SPM level 2 of a query image.

similarity with Cosine Distance between query feature $\hat{Q}^i$ and each patch feature (visual word feature) $\hat{P}^i_j$ in this i[th] page of the manuscript.

$$d = 1 - \frac{\hat{Q}^i . \hat{P}^i_j}{\left\|\hat{Q}^i\right\| \left\|\hat{P}^i_j\right\|} \tag{10}$$

2) Selection of N smallest distance to build the map of spotting area: For each query image, we selected N patches with the smallest distance between patch feature and query feature. In our experiments, we tested the value of N=75,100,125. All patches selected are affixed on their position to build the map of spotting area (Figure 41).



Figure 41. Map of spotting area of all selected patches.

To filter the best-selected patches from the previous step, we can apply the Longest Common Subsequence (LCS) (cor, 2001) or Longest Weighted Profile (LWP) technique

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 265 of 404

(Dovgalecs et al., 2013). To perform the LCS and LWP algorithm, the visual word of the query feature and patch feature should be constructed by concatenating each row of the visual word into one-dimensional row vector. In the final step, to ignore redundant and overlapping patches, we propose and apply a simple patch selection algorithm.

4.5. Longest Common Subsequence

Longest Common Subsequence (LCS) technique is applied to measure spatial common subsequence between selected patch feature and query feature. The common subsequence must appear in the same order, but not necessarily consecutively (cor, 2001). Figure 42 shows the scheme of the LCS algorithm and Algorithm 1 describes the pseudocode of the LCS algorithm. We computed the length of LCS by using a matrix S. The elements of matrix S will be computed in row-major order, started from the first row, from left to right. The element Sij depends only on whether sequence element $Q_i = Y_j$ and the values of element $S_{i-1,j}$, $S_{i,j-1}$, and $S_{i-1,j-1}$ which are computed before element S_{ij}. The last S_{ij} contains the length of LCS between sequence Q and Y. We divided the length of LCS with the minimum length between two input sequences. If the common subsequence value is greater than a threshold value (T), the selected patch is defined as the selected spotting area (Figure 43). Based on our experiments, the threshold value is empirically tested and set with T=0.35,0.40.

	Yj	A	D	C	B	E
Xj	0	0	0	0	0	0
A	0	1	1	1	1	1
B	0	1	1	1	2	2
E	0	1	1	1	1	3
C	0	1	1	2	2	3
C	0	1	1	2	2	3
D	0	1	2	2	2	3
E	0	1	2	2	2	3

Figure 42. LCS technique.



Figure 43. The selected patches after LCS technique of Figure 41.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

input : Q - query sequence composed of m visual words
Y - test sequence composed of n visual words
output: sQY - similarity score
begin
LQ := length(Q)
LY := length(Y)
S := array $(0...m, 0...n) \leftarrow 0$
for $i := 1$ *to* m **do**
for $j := 1$ *to* n **do**
if $Q_i = Y_j$ **then**
$S_{ij} := S_{i-1,j-1} + 1$
end
else
$S_{ij} := max(S_{i,j-1}, S_{i-1,j})$
end
end
end
$S_{mn}/min(LQ, LY))$
end

Algorithm 1: LCS algorithm.

4.6. Longest Weighted Profile

The Longest Weighted Profile (LWP) algorithm was proposed in (Dovgalecs et al., 2013). The LWP algorithm tries to eliminate false positives without losing true ones by counting not only the strict match and mismatch, but by tolerating the small random variations between cluster centers of visual word. This information is encoded in a symmetric similarity matrix. This algorithm takes two inputs of visual word sequences, Q and Y, and an inter-cluster similarity matrix M. Matrix M describes the similarity between two cluster centers. Matrix M can be computed using formula as follows (Dovgalecs et al., 2013).

$$M_{i,j} = \max\left(0, \frac{\langle \mu_i, \mu_j \rangle}{\| \mu_i \| \| \mu_j \|}\right)^{\tau}, \forall_{i,j} \in \{1, ..., K\} \tag{11}$$

Where are the concatenated SIFT feature space cluster centers, k is the number of clusters, and τ ¿0. As in (Dovgalecs et al., 2013), we used τ = 50 in our experiments. Algorithm 2 describes the pseudocode of the LWP algorithm. If , the matrix M will become an identity matrix and LWP algorithm reduces to the same LCS algorithm. As in the experiment with LCS, to filter the spotting areas, we empirically set the threshold value T=0.35,0.40.

4.7. Patch Selection Algorithm

For final selection of spotting area based on map of selected patches, to ignore some redundant and overlapping patches, we proposed and applied a simple patch selection algorithm to locate the final spotting area on document image (Figure 44). For a non-overlapping

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Cerence Operating Company, Ex. 2014
Page 267 of 404

patch area, a single patch is directly selected as a final spotting area. In a group of overlapping patches, we calculated the number of overlapping patches on each pixel in this area, and we choose all pixels which contain the maximum number of overlapping patches as a center of a new spotting area. A new spotting area is defined as a minimum rectangle area to cover all those pixels. If this new spotting area is smaller than the size of query image, it will be adjusted to the size of the query image.

input : Q - query sequence composed of m visual words
Y - test sequence composed of n visual words
M - k x k inter-cluster similarity matrix
output: sQY - similarity score
begin
LQ := length(Q)
LY := length(Y)
S := array $(0...m, 0...n) \leftarrow 0$
for $i := 1$ *to* m **do**
for $j := 1$ *to* n **do**
if $Q_i = Y_j$ **then**
$S_{ij} := S_{i-1,j-1} + 1$
end
else
$\wedge_{ij} := S_{i-1,j-1} + M_{Qi,Yj}$ $S_{ij} := max(S_{i,j-1}, S_{i-1,j,\wedge_{ij}})$
end
end
end
$S_{mn}/min(LQ, LY)$
end

Algorithm 2: LWP algorithm (adapted from Dovgalecs et al. (2013)).



Figure 44. Spotting Area after patch selection algorithm of Figure 43.

5. Corpus and Dataset

5.1. The Palm Leaf Manuscript Corpus and the Digitization Process

The first corpus of palm leaf manuscript images which was collected from Southeast Asia are the sample images of the palm leaf manuscripts from Bali, Indonesia (Kesiman et al., 2015a,b). In order to obtain the variety of the manuscript images (different content and writer), the sample images were collected from 23 different collections (contents), which come from 5 different locations (regions). From those 23 collections, 393 pages of palm leaf manuscript were captured. A summary of the collections is listed in Table 5.

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267 Exhibit 1011 Ex. 2011
Complimentary Contributor Copy
Page 268 of 404

To capture the manuscript images, a Canon EOS 5D Mark III camera was used. The camera settings are as follows (Kesiman et al., 2015b): F-stop: f/22 (diaphragm), exposure time: 1/50 sec, ISO speed: ISO-6400, focal length: 70 mm, flash: On - 1/64, distance to object: 76 cm, focus: Quick mode - Auto selection 'On'. A black box camera support by wood was also used to avoid the irregular lighting/luminance condition and to fits the semi-outdoor capturing location (Figure 45). This camera support was optimally designed to be used under some restricted conditions given by the museum or the owner of the manuscripts. Two additional lights have been added inside the black box support. These lights are a White Neon 50 cm of 20 watts. Thumbnail samples of the captured images are showed in Figure 46.

5.2. The Dataset of AMADI_LontarSet

In order to develop and to evaluate the performance of the document analysis methods, the dataset and the corresponding ground truth data are required. Therefore, creating a new dataset and ground truth image for palm leaf manuscripts was a necessary step for the research community. Under the scheme of the AMADI (Ancient Manuscripts Digitization and Indexation) Project, we have built the AMADI_LontarSet (Kesiman et al., 2016b), the first handwritten Balinese palm leaf manuscript dataset. It includes three components of the dataset as follows: binarized images ground truth dataset, word annotated images dataset, and isolated character annotated images dataset. The resume of the dataset is presented on Table 6. The whole dataset is publicly available for scientific use on http://amadi.univ-lr.fr/ICFHR2016_Contest/ (Kesiman et al., 2016b).

5.2.1. The Binarized Images Ground Truth Dataset

Table 7 shows the summary of binarized images ground truth dataset for the AMADI_LontarSet (Kesiman et al., 2016b). For the training-based binarization method, we divide our dataset into two subsets: 50 images for training and 50 images for testing. Figure 47 shows some samples of binarized images ground truth from our dataset. For more detail about the analysis of ground truth binarized image variability of palm leaf manuscripts, please refer to the previous section "Analysis of ground truth binarized image variability".

5.2.2. The Word Annotated Images Dataset

To create the word annotated ground truth dataset of the manuscript, we asked a collaborative work between the Balinese script philologists, students from the Department of Informatics, and students from the Department of Balinese Literature. The philologists read the manuscripts and create the Latin transcription. Based on this Latin transcription, a pair of students (one student in Informatics and one student in Balinese Literature) works together to segment and to annotate each word in manuscripts. The validation and correction of word annotation are done based on the expertise of the philologists. Any further discussion remains open between the philologists and the ground truthers to correct and to validate the transcription while the annotation process.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 269 of 404

Table 5. Corpus of palm leaf manuscripts from Bali, Indonesia (Kesiman et al., 2016b).

Location	Collection Code	Content	Nb of captured pages
Museum Gedong Kertya, Singaraja (10 collections)	IIA-10-1534	*Awig-awig Desa Tunju*	8
	IIA-5-789	*Sima Desa Tejakula*	8
	IIB-2-180	*Dewa Sasana*	8
	IIIB-12-306	*Panugrahan Bhatara Ring Pura Pulaki*	8
	IIIB-42-1526	*Buwana*	8
	IIIB-45-2296	*Pambadah*	8
	IIIC-19-1293	*Krakah Sang Graha*	8
	IIIC-20-1397	*Taru Pramana*	8
	IIIC-23-1506	*Siwa Kreket*	8
	IIIC-24-1641	*Tikas Patanganan Weda*	8
Museum Bali, Denpasar (4 collections)	MB-AdiParwa(Purana)-5338.2-IV.a	*Adi Parwa (Purana)*	40
	MB-AjiGriguh-5783-107.2	*Aji Griguh*	20
	MB-ArjunaWiwaha-GrantangBasaII	*Arjuna Wiwaha-Grantang Basa II*	30
	MB-TaruPramana	*Taru Pramana*	40
Village of Jagaraga, Buleleng (7 collections)	JG-01	*Unknown*	16
	JG-02	*Unknown*	10
	JG-03	*Unknown*	16
	JG-04	*Unknown*	12
	JG-05	*Unknown*	8
	JG-06	*Unknown*	5
	JG-07	*Unknown*	10
Village of Susut, Bangli (1 collection)	Bangli	*Sabung Ayam*	82
Village of Rendang, Karangasem (1 collection)	WN	*Surat Jual Beli Tanah*	24
TOTAL			393

Table 6. Global summary of dataset of palm leaf manuscript images

No.	Data collection	Format	Quantity
1.	Original Images	RGB Color image - JPG	$\pm$ 300 images
2.	Transcription of manuscripts	TXT	$\pm$ 300 text
3.	Binarized ground truth image Version 1	Binary image - BMP	100 images
4.	Binarized ground truth image Version 2	Binary image - BMP	100 images
5.	Word annotated segment images	RGB Color image - JPG	$\pm$ 34,520 images from $\pm$ 8,724 unique words
6.	Character annotated segment images	RGB Color image - JPG	$\pm$ 27,496 images from $\pm$ 133 classes of character

Table 7. Summary of binarized images ground truth dataset for the AMADI_LontarSet (Kesiman et al., 2016b)

No.	Data	Format	Qty.
1.	Original Images of Manuscript	RGB Color image - JPG	100 images
2.	Binarized Ground Truth Image (1stground truther)	Binary image - BMP	100 images
3.	Binarized Ground Truth Image (2ndground truther)	Binary image - BMP	100 images

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 270 of 404

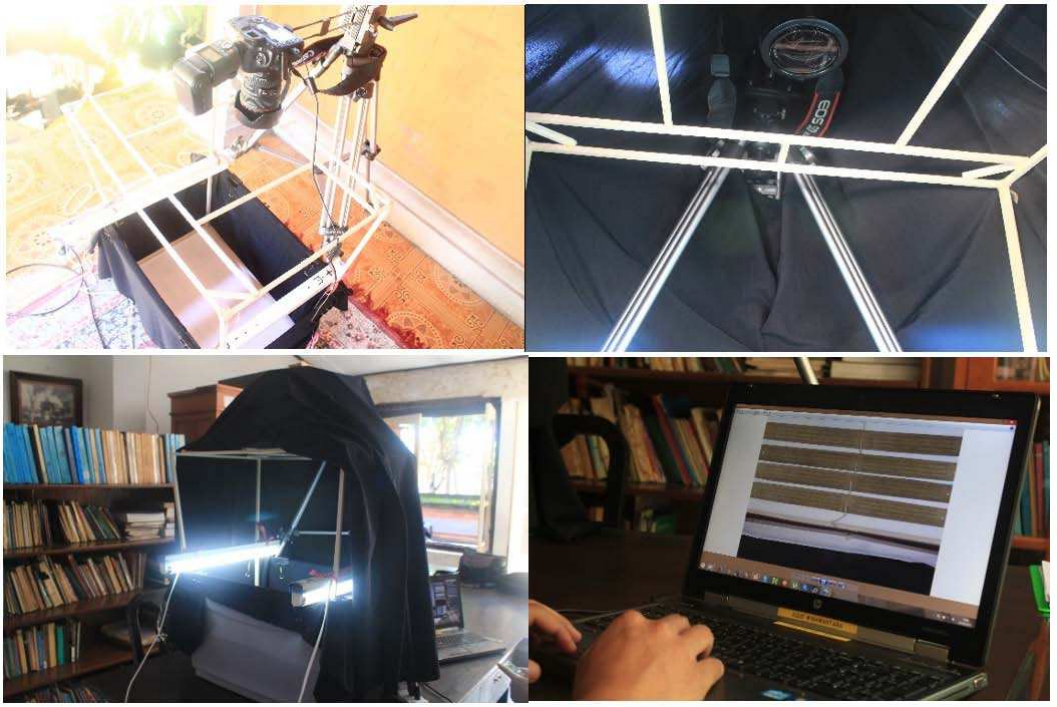


Figure 45. Camera support for digitizing process of palm leaf manuscripts.

We used ALETHEIA[9], an advanced document layout and text ground-truthing system (Chamchong and Fung, 2011), to segment and to annotate the words (Figure 48). After the segmentation and the annotation process, the manuscript images are then cropped based on word polygon coordinates in the XML file produced by ALETHEIA (Figure 49). Table 8 shows the summary of word annotated images dataset for the AMADI_LontarSet (Kesiman et al., 2016b).

Table 8. Summary of word annotated images dataset for the AMADI_LontarSet (Kesiman et al., 2016b)

No.	Data	Format	Qty.
1.	Training Set: Original Images of Manuscript	RGB Color image - JPG	130 images
2.	Training Set: Transcription of manuscript of No 1	TXT	130 text files
3.	Training Set: Word annotated images of No 1	RGB Color image - JPG	15,022 images
4.	Testing Set: Original Images of Manuscript	RGB Color image - JPG	100 images
5.	Testing Set: Transcription of manuscript of No 4	TXT	100 text files
6.	Testing Set: Word annotated images of No 4	RGB Color image - JPG	10,475 images
7.	Testing Set: Selected word annotated images as query-by-example	RGB Color image - JPG	36 images
8.	Testing Set: Ground truth images for all query images of No 7	RGB Color image - JPG	257 images

[9]http://www.primaresearch.org/tools/Aletheia

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 271 of 404

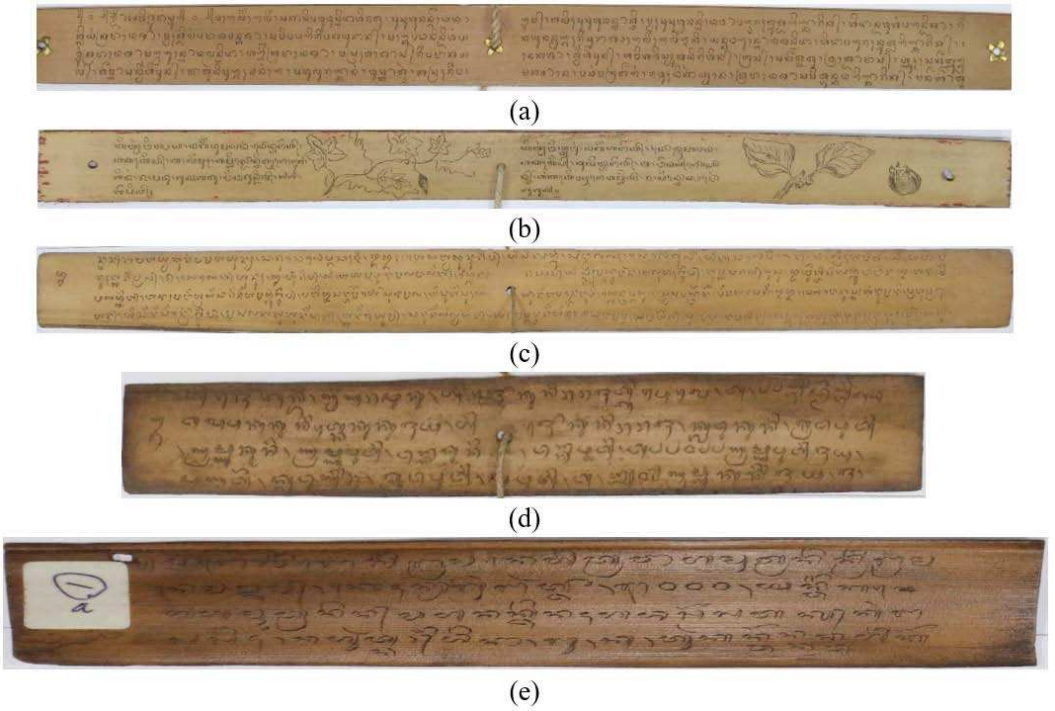


Figure 46. Sample images of palm leaf manuscript from a) Museum Gedong Kertya, Singaraja, b) Museum Bali, Denpasar, c) Village of Jagaraga, Buleleng, d) Village of Susut, Bangli, e) Village of Rendang, Karangasem (Kesiman et al., 2016b).

5.2.3. The Isolated Character Annotated Images Dataset

By using the collection of word annotated images which were produced in our previous ground truthing process, we collected our isolated handwritten Balinese character dataset. First, we applied Otsu (Pratikakis et al., 2013; Messaoud et al., 2011) binarization method to all word patch images. We automatically extracted all connected component found on the binarized word patch images. Our Balinese philologists then annotated manually all connected components that represent a correct character in Balinese script. To facilitate the work of the philologists, we developed a simple web-based user interface for this character annotation process (Figure 50). With this web-based interface, more than one philologist can work together to verify, to correct and to validate the annotation of the characters. All annotated characters are displayed based on their given class. A hyperlink from each annotated character to their corresponding word annotated images is provided to allow the philologists to verify and to correct the annotation (Figure 51).

All patch images that have been segmented and annotated at character level constitute the isolated character dataset. Table 9 shows the summary of isolated character annotated images dataset for the AMADI_LontarSet (Kesiman et al., 2016b). The number of sample images for each class is different. Some classes are frequently found in our collection of palm leaf manuscripts, but some others are rarely used. Thumbnail samples of these character annotated images are showed in Figure 52.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 272 of 404

Table 9. Summary of binarized images ground truth dataset for the amadi_lontarset (Kesiman et al., 2016b)

No.	Data	Format	Qty.
1.	Training Set: Character annotated images	RGB Color image - JPG	133 classes - 11,710 images
2.	Testing Set: Character annotated images	RGB Color image - JPG	133 classes - 7,673 images

In a near future, we plan to develop the dataset in term of data quantity and variety to be able to provide sufficiently a larger train data set for document analysis methods.

6. Experiments

6.1. Experiment on Isolated Character Recognition

We present this experimental study on feature extraction methods for character recognition of Balinese script on palm leaf manuscripts (Kesiman et al., 2016c). We investigated and evaluated the performance of 10 feature extraction methods and the proposed combination of features in 29 different schemes. For all experiments, a set of image patches containing Balinese characters from the original manuscripts will be used as input, and a correct class of each character should be identified as a result. We used k=5 for the k-NN classifier, and all images are resized to 50x50 pixels (the approximate average size of a character in the collection), except for Gradient features where images are resized to 81x81 pixels to get evenly 81 blocks of 9x9 pixels, as described in (Fujisawa et al., 1999).

The results (Table 10) show that the recognition rate of NPW features can be significantly increased (up to 10%) by applying it on the four directional Kirsch edge images (NPW-Kirsch method). Then, by combining this NPW-Kirsch features, HoG features, and Zoning method can increase the recognition rate up to 85% [6]. In these experiments, the number of training dataset for each class is not balanced. But this condition was already clearly stated and can not be avoided in our case of IHCR development for Balinese script on palm leaf manuscripts. Some ancient characters are not frequently found in our collection of palm leaf manuscripts.

6.2. Experiment on Word Spotting

In this experiment, we evaluated the performance of word spotting method with Bag of Visual Word (BoVW) in six different frameworks (Figure 53). We calculated the mean Recall (mR) and the mean average Precision (maP) (Rusinol et al., 2011; Rothacker et al., 2013a,b) of spotting area based on ground truth word-level annotated patch images of the testing subset (Table 11). A spotting area is considered as relevant if it overlaps more than 50% of a ground truth word-level patch area containing the same query word (Dovgalecs et al., 2013; Rusinol et al., 2011; Rothacker et al., 2013a) and if the size of the spotting area (width and height) is not twice bigger than the size of ground truth area.

We can see in Table 11 that the mean recall value and the mean average precision is high enough for the framework with LSI technique combined with LCS or LWP technique. In general, the decreasing number of selected patches (N) can increase the mean average

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 273 of 404

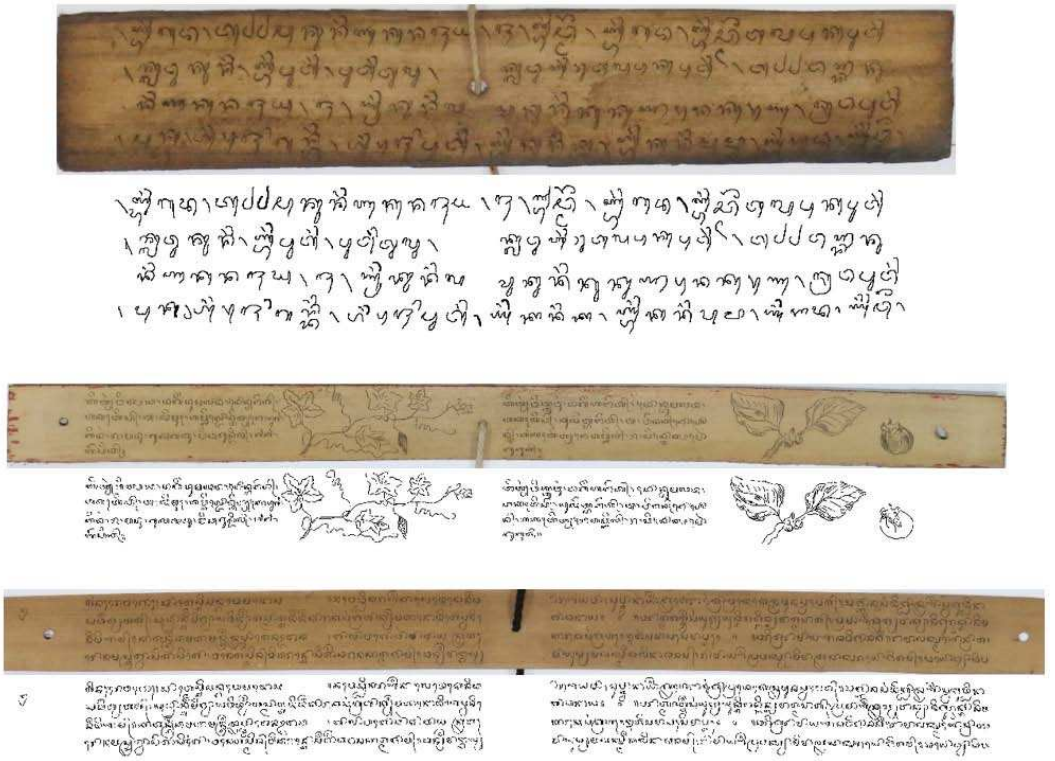


Figure 47. Samples of binarized images ground truth dataset (Kesiman et al., 2016b).

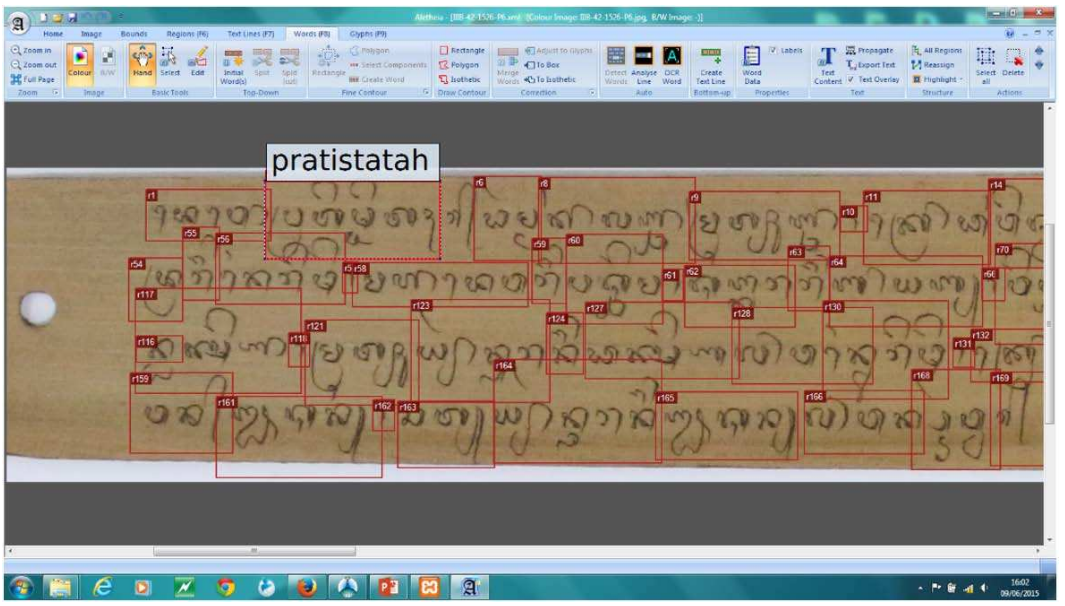


Figure 48. Word annotation with ALETHEIA (Kesiman et al., 2016b).

precision value. This is because, in the collection of palm leaf manuscripts, a specific word

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 274 of 404



Figure 49. Samples of word annotated images (Kesiman et al., 2016b).



Figure 50. Screenshot of web based user interface for the character annotation process.

is normally found only on a very limited number of pages. Most of the patches with small distance feature with query image are normally already found in one page. By limiting the number of selected patches, it can decrease the number of spotted area in the other pages of manuscript which do not contain the query word. LCS and LWP technique increase the mean average precision value.

Conclusions and Future Work

This chapter described in detail the historical handwritten document analysis for Southeast Asian palm leaf manuscripts by reporting the latest findings in studies and experimental results of document analysis tasks which range from corpus collection, ground truth data generation, binarization process to the isolated character recognition and the word spotting tasks. For the degraded ancient document image analysis, the choice of ground truth data set and the variability within the ground truth should be analyzed quantitatively before the

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 275 of 404



Figure 51. Screenshot of character class verification.



Figure 52. Samples of character-level annotated patch images of Balinese script on palm leaf manuscripts (Kesiman et al., 2016b).

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 276 of 404

Table 10. Recognition rate from all schemes of experiment (Kesiman et al., 2016c)

No.	Method	Feature Dim.	Classifier	Recog. Rate %
1.	Histogram Projection (Binary)	100	SVM	26,313
2.	Celled Projection (Binary)	500	SVM	49,9414
3.	Celled Projection (Binary)	500	k-NN	76,1632
4.	Distance Profile (Binary)	200	SVM	40,1277
5.	Distance Profile (Binary)	200	k-NN	58,947
6.	Distance Profile (Skeleton)	200	SVM	36,7653
7.	Crossing (Binary)	100	SVM	15,0007
8.	Zoning (Binary)	205	SVM	50,6451
9.	Zoning (Binary)	205	k-NN	78,5351
10.	Zoning (Skeleton)	205	SVM	41,848
11.	Zoning (Grayscale)	205	SVM	52,4176
12.	Zoning (Grayscale)	205	k-NN	66,128
13.	Gradient Feature (Gray)	400	SVM	60,0417
14.	Gradient Feature (Gray)	400	k-NN	72,5792
15.	Moment Hu (Gray)	56	SVM	33,481
16.	Moment Hu (Gray)	56	k-NN	33,481
17.	HoG (Gray)	1984	SVM	71,2759
18.	HoG (Gray)	1984	k-NN	84,3477
19.	NPW (Binary)	100	SVM	51,388
20.	NPW (Gray)	100	SVM	54,1249
21.	Kirsch (Gray)	100	SVM	62,4528
22.	HoG with Zoning (Gray)	1984	SVM	69,6859
23.	HoG with Zoning (Gray)	1984	k-NN	83,5006
24.	NPW-Kirsch (Gray)	400	SVM	63,5736
25.	NPW-Kirsch (Gray)	400	k-NN	76,7105
26.	HoG on Kirsch edge (Gray)	1984*4	k-NN	82,0931
27.	HoG + NPW-Kirsch (Gray)	1984+400	k-NN	84,7517
28.	Zoning + Celled Projection (Binary)	205+500	k-NN	77,701
29.	**HoG + NPW-Kirsch (Gray) + Zoning (Binary)**	**1984+400+205**	**k-NN**	**85,1557**

performance measure of any binarization methods. In the case of a manuscript with specific ancient characters, the qualitative observation and validation should also be made by the philologist to guarantee the correctness of the binarized characters on the manuscripts. A proper and robust combination of feature extraction methods can increase the character recognition rate. This study shows that the recognition rate of isolated character recognition of Balinese script can be significantly increased by applying NPW features on four directional Kirsch edge images. And the use of NPW on Kirsch features in combination with HoG features and Zoning method can increase the recognition rate up to 85%. The results of the study on word spotting show the challenging characteristics of a manuscript collection

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 277 of 404

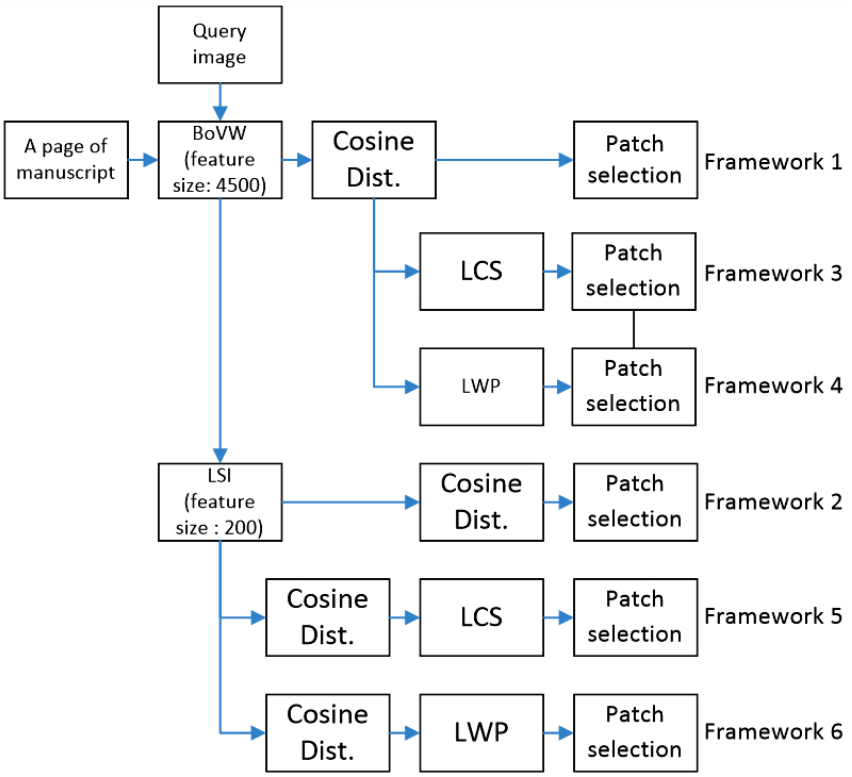


Figure 53. Framework diagram of experiments.

with single script and multi-writer. Even though the methods and frameworks for this query based word spotting technique are normally evaluated only on a single writer manuscript collection, the results of this experiment show that the powerful framework combination of BoVW with LSI, LCS and LWP can still provide a possibility to support the indexing and word spotting system for multi-writer palm leaf manuscript images.

We have built the AMADI_LontarSet, the first handwritten Balinese palm leaf manuscript dataset. It includes three components of dataset as follows: binarized images ground truth dataset, word annotated images dataset, and isolated character annotated images dataset. To improve the accuracy of character and text recognition of Balinese script on palm leaf manuscripts, a lexicon-based statistical approach is needed. The lexicon dataset will provide useful information about textual correlation of Balinese script (between characters, syllables, and words). This information will be needed in the correction step of text recognition when the physical feature description is failed to do the recognition process. Our future interests are: - to build an optimal lexicon dataset for our system, in term of quantity and completeness of the dataset. - to define an appropriate lexicon information (characters, syllables and words level) for Balinese script. Writing in Balinese script, there is no space between words in a text line. Most of the text recognition methods which naturally proposed a sequential process to recognize the words as entity/unit will face this characteristic as a very challenging task. The data representation of all words in some specific compositions of part-of-words (POW) can feed the recognizer with useful contextual

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 278 of 404

Table 11. Results of experiment.

		Param		Recall			maP		
No.	**Framework**	**N**	**T**	**min**	**max**	**mR**	**min**	**max**	**maP**
1	BoVW	75	-	0	100	12,48	0	100	25,19
2	BoVW	100	-	0	100	29,6	0	100	34,23
3	BoVW	125	-	0	100	23,43	0	100	30,2
4	BoVW+LSI	75	-	0	100	28,57	0	100	22,51
5	BoVW+LSI	100	-	0	100	29,01	0	100	26,27
6	BoVW+LSI	125	-	0	100	28,4	0	100	24,81
7	BoVW+LCS	75	35	0	100	33,27	0	100	39,75
8	BoVW+LCS	100	35	0	100	33,2	0	100	38,03
9	BoVW+LCS	125	35	0	100	33,19	0	100	37,53
10	BoVW+LCS	75	40	0	100	34,02	0	100	37,67
11	BoVW+LCS	100	40	0	100	34	0	100	34,8
12	BoVW+LCS	125	40	0	100	33,29	0	100	34,23
13	BoVW+LWP	75	35	0	100	35,01	0	100	39,41
14	BoVW+LWP	100	35	0	100	34,64	0	100	37,45
15	BoVW+LWP	125	35	0	100	35,55	0	100	35,82
16	BoVW+LWP	75	40	0	100	35,04	0	100	37,9
17	BoVW+LWP	100	40	0	100	34,12	0	100	35,09
18	BoVW+LWP	125	40	0	100	33,42	0	100	34,86
19	BoVW+LSI+LCS	75	35	0	100	32,6	0	100	33,71
20	BoVW+LSI+LCS	100	35	0	100	33,5	0	100	30,41
21	BoVW+LSI+LCS	125	35	0	100	33,14	0	100	26,58
22	BoVW+LSI+LCS	75	40	0	100	33,21	0	100	31,39
23	BoVW+LSI+LCS	100	40	0	100	33,6	0	100	28,13
24	BoVW+LSI+LCS	125	40	0	100	34,03	0	100	24,42
25	BoVW+LSI+LWP	75	35	0	100	30,4	0	70	18,09
26	BoVW+LSI+LWP	100	35	0	100	35,04	0	100	30,35
27	BoVW+LSI+LWP	125	35	0	100	35,32	0	100	26,94
28	BoVW+LSI+LWP	75	40	0	100	33,43	0	55,56	17,74
29	BoVW+LSI+LWP	100	40	0	100	33,57	0	100	28,66
30	BoVW+LSI+LWP	125	40	0	100	34,03	0	100	24,42

knowledge. The multiword expression/unit (MWE/MWU) will be needed to model these contextual information from the manuscript corpus. The relation between the words and their corresponding multiword expression/unit models can help the text recognition system to do the post processing task, such as correction and validation step of the recognized texts. To support our project, we are really interested in building and constructing such lexicon model based on the multiword expression/unit. Based on our knowledge, there is no ready-to-use lexicon database for Balinese-Kawi language which was used in our manuscript corpus. We plan to define and to construct a sufficient and an optimal lexicon model from our character-level and word-level annotated data.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 279 of 404

Acknowledgment

The authors would like to thank Museum Gedong Kertya, Museum Bali, and all families in Bali, Indonesia, for providing us the samples of palm leaf manuscripts, and the students from the Department of Informatics Education and the Department of Balinese Literature, Ganesha University of Education for helping us in ground truthing process for this research project. This work is also supported by the DIKTI BPPLN Indonesian Scholarship Program and the STIC Asia Program implemented by the French Ministry of Foreign Affairs and International Development (MAEDI).

References

(2001). Introduction to algorithms, 2nd ed.

Aggarwal, A., Singh, K., and Singh, K. (2015). Use of gradient technique for extracting features from handwritten gurmukhi characters and numerals. *Procedia Computer Science*, 46:1716–1723.

Almeida, J., Torres, R. d. S., and Goldenstein, S. (2009). Sift applied to cbir. *Revista de Sistemas de Informacao da FSMA n*, 4:41–48.

Arica, N. and Yarman-Vural, F. T. (2002). Optical character recognition for cursive handwriting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(6):801–813.

Ashlin Deepa, R. and Rajeswara Rao, R. (2014). Feature extraction techniques for recognition of malayalam handwritten characters: Review. *International Journal of Advanced Trends in Computer Science and Engineering*, 3(1):481–485.

Auclair, A., Cohen, L. D., and Vincent, N. (2007). How to use sift vectors to analyze an image with database templates. In *International Workshop on Adaptive Multimedia Retrieval*, pages 224–236. Springer.

Bal, G., Agam, G., Frieder, O., and Frieder, G. (2008). Interactive degraded document enhancement and ground truth generation. In *Electronic Imaging 2008*, pages 68150Z–68150Z. International Society for Optics and Photonics.

Blumenstein, M., Verma, B., and Basli, H. (2003). A novel feature extraction technique for the recognition of segmented handwritten characters. In *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*, pages 137–141. IEEE.

Bokser, M. (1992). Omnidocument technologies. *Proceedings of the IEEE*, 80(7):1066–1078.

Burie, J.-C., Coustaty, M., Hadi, S., Kesiman, M. W. A., Ogier, J.-M., Paulus, E., Sok, K., Sunarya, I. M. G., and Valy, D. (2016). Icfhr 2016 competition on the analysis of handwritten text in images of balinese palm leaf manuscripts. In *15th International Conference on Frontiers in Handwriting Recognition 2016*, pages 596–601.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 280 of 404

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698.

Chamchong, R. and Fung, C. C. (2011). Character segmentation from ancient palm leaf manuscripts in thailand. In *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, pages 140–145. ACM.

Chamchong, R. and Fung, C. C. (2012). Text line extraction using adaptive partial projection for palm leaf manuscripts from thailand. In *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*, pages 588–593. IEEE.

Chamchong, R., Fung, C. C., and Wong, K. W. (2010). Comparing binarisation techniques for the processing of ancient manuscripts. In *Cultural Computing*, pages 55–64. Springer.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Dovgalecs, V., Burnett, A., Tranouez, P., Nicolas, S., and Heutte, L. (2013). Spot it! finding words and patterns in historical documents. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 1039–1043. IEEE.

Feng, M.-L. and Tan, Y.-P. (2004). Contrast adaptive binarization of low quality document images. *IEICE Electronics Express*, 1(16):501–506.

Fischer, A., Keller, A., Frinken, V., and Bunke, H. (2012). Lexicon-free handwritten word spotting using character hmms. *Pattern Recognition Letters*, 33(7):934–942.

Fujisawa, Y., Shi, M., Wakabayashi, T., and Kimura, F. (1999). Handwritten numeral recognition using gradient and curvature of gray scale image. In *Document Analysis and Recognition, 1999. ICDAR'99. Proceedings of the Fifth International Conference on*, pages 277–280. IEEE.

Fung, C. C. and Chamchong, R. (2010). A review of evaluation of optimal binarization technique for character segmentation in historical manuscripts. In *Knowledge Discovery and Data Mining, 2010. WKDD'10. Third International Conference on*, pages 236–240. IEEE.

Gatos, B., Ntirogiannis, K., and Pratikakis, I. (2011). Dibco 2009: document image binarization contest. *International Journal on Document Analysis and Recognition (IJDAR)*, 14(1):35–44.

Govindan, V. and Shivaprasad, A. (1990). Character recognition—a review. *Pattern recognition*, 23(7):671–683.

Gupta, M. R., Jacobson, N. P., and Garcia, E. K. (2007). Ocr binarization and image pre-processing for searching historical documents. *Pattern Recognition*, 40(2):389–397.

He, J., Do, Q., Downton, A. C., and Kim, J. (2005). A comparison of binarization methods for historical archive documents. In *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*, pages 538–542. IEEE.

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 281 of 404

Hossain, M. Z., Amin, M. A., and Yan, H. (2012). Rapid feature extraction for optical character recognition. *arXiv preprint arXiv:1206.0238*.

Howe, N. R. (2013). Document binarization with automatic parameter tuning. *International Journal on Document Analysis and Recognition (IJDAR)*, 16(3):247–258.

Jin, Z., Qi, K., Zhou, Y., Chen, K., Chen, J., and Guan, H. (2009). Ssift: An improved sift descriptor for chinese character recognition in complex images. In *Computer Network and Multimedia Technology, 2009. CNMT 2009. International Symposium on*, pages 1–5. IEEE.

Kesiman, M. W. A., Burie, J.-C., and Ogier, J.-M. (2016a). A new scheme for text line and character segmentation from gray scale images of palm leaf manuscript. In *15th International Conference on Frontiers in Handwriting Recognition 2016, At Shenzhen, China*, pages 325–330.

Kesiman, M. W. A., Burie, J.-C., Ogier, J.-M., Wibawantara, G. N. M. A., and Sunarya, I. M. G. (2016b). Amadi_lontarset: The first handwritten balinese palm leaf manuscripts dataset. In *15th International Conference on Frontiers in Handwriting Recognition 2016*, pages 168–172.

Kesiman, M. W. A., Prum, S., Burie, J.-C., and Ogier, J.-M. (2015a). An initial study on the construction of ground truth binarized images of ancient palm leaf manuscripts. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*.

Kesiman, M. W. A., Prum, S., Burie, J.-C., and Ogier, J.-M. (2016c). Study on feature extraction methods for character recognition of balinese script on palm leaf manuscript images. In *23rd International Conference on Pattern Recognition*, pages 4006–4011.

Kesiman, M. W. A., Prum, S., Sunarya, I. M. G., Burie, J.-C., and Ogier, J.-M. (2015b). An analysis of ground truth binarized image variability of palm leaf manuscripts. In *Image Processing Theory, Tools and Applications (IPTA), 2015 International Conference on*, pages 229–233. IEEE.

Kesiman, M. W. A., Valy, D., Burie, J.-C., Paulus, E., Sunarya, I. M. G., Hadi, S., Sok, K. H., and Ogier, J.-M. (2017). Southeast asian palm leaf manuscript images: a review of handwritten text line segmentation methods and new challenges. *Journal of Electronic Imaging*, 26(1):011011–011011.

Khayyat, M., Lam, L., and Suen, C. Y. (2013). Verification of hierarchical classifier results for handwritten arabic word spotting. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 572–576. IEEE.

Khurshid, K., Siddiqi, I., Faure, C., and Vincent, N. (2009). Comparison of niblack inspired binarization methods for ancient documents. In *IS&T/SPIE Electronic Imaging*, pages 72470U–72470U. International Society for Optics and Photonics.

Kumar, S. (2010). Neighborhood pixels weights-a new feature extractor. *International Journal of Computer Theory and Engineering*, 2(1):69.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 282 of 404

Kumar, S. (2011). Study of features for hand-printed recognition. *Int. J. Comput. Electr. Autom. Control Inf. Eng. 5*.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, pages 2169–2178. IEEE.

Ledwich, L. and Williams, S. (2004). Reduced sift features for image retrieval and indoor localisation. In *Australian conference on robotics and automation*, volume 322, page 3. Citeseer.

Lee, D.-R., Hong, W., and Oh, I.-S. (2012a). Segmentation-free word spotting using sift. In *Image Analysis and Interpretation (SSIAI), 2012 IEEE Southwest Symposium on*, pages 65–68. IEEE.

Lee, D.-R., Hong, W., and Oh, I.-S. (2012b). Segmentation-free word spotting using sift. In *Image Analysis and Interpretation (SSIAI), 2012 IEEE Southwest Symposium on*, pages 65–68. IEEE.

Lehal, G. S. and Singh, C. (2000). A gurmukhi script recognition system. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 2, pages 557–560. IEEE.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Messaoud, I. B., El Abed, H., Märgner, V., and Amiri, H. (2011). A design of a preprocessing framework for large database of historical documents. In *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, pages 177–183. ACM.

Nafchi, H. Z., Ayatollahi, S. M., Moghaddam, R. F., and Cheriet, M. (2013). An efficient ground truthing tool for binarization of historical manuscripts. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 807–811. IEEE.

Neha J. Pithadia, D. V. D. N. (2015). A review on feature extraction techniques for optical character recognition. *Int. J. Innov. Res. Comput. Commun. Eng. 3*.

Ntirogiannis, K., Gatos, B., and Pratikakis, I. (2008). An objective evaluation methodology for document image binarization techniques. In *Document Analysis Systems, 2008. DAS'08. The Eighth IAPR International Workshop on*, pages 217–224. IEEE.

Ntirogiannis, K., Gatos, B., and Pratikakis, I. (2013). Performance evaluation methodology for historical document image binarization. *IEEE Transactions on Image Processing*, 22(2):595–609.

Pal, U. and Chaudhuri, B. (2004). Indian script character recognition: a survey. *Pattern Recognit.*, 37(9):1887–1899.

Pal, U., Jayadevan, R., and Sharma, N. (2012). Handwriting recognition in indian regional scripts: a survey of offline techniques. *ACM Transactions on Asian Language Information Processing (TALIP)*, 11(1):1.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 283 of 404

Pratikakis, I., Gatos, B., and Ntirogiannis, K. (2013). Icdar 2013 document image binarization contest (dibco 2013). In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 1471–1476. IEEE.

Rais, N. B., Hanif, M. S., and Taj, I. A. (2004). Adaptive thresholding technique for document image analysis. In *Multitopic Conference, 2004. Proceedings of INMIC 2004. 8th International*, pages 61–66. IEEE.

Ramteke, R. (2010). Invariant moments based feature extraction for handwritten devanagari vowels recognition. *International Journal of Computer Applications*, 1(18):1–5.

Rani, M. and Meena, Y. K. (2011). An efficient feature extraction method for handwritten character recognition. In *International Conference on Swarm, Evolutionary, and Memetic Computing*, pages 302–309. Springer.

Rothacker, L., Fink, G. A., Banerjee, P., Bhattacharya, U., and Chaudhuri, B. B. (2013a). Bag-of-features hmms for segmentation-free bangla word spotting. In *Proceedings of the 4th International Workshop on Multilingual OCR*, page 5. ACM.

Rothacker, L., Rusinol, M., and Fink, G. A. (2013b). Bag-of-features hmms for segmentation-free word spotting in handwritten documents. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 1305–1309. IEEE.

Rusinol, M., Aldavert, D., Toledo, R., and Llados, J. (2011). Browsing heterogeneous document collections by a segmentation-free word spotting method. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 63–67. IEEE.

Saund, E., Lin, J., and Sarkar, P. (2009). Pixlabeler: User interface for pixel-level labeling of elements in document images. In *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*, pages 646–650. IEEE.

Sauvola, J. and Pietikäinen, M. (2000). Adaptive document image binarization. *Pattern recognition*, 33(2):225–236.

Sharma, D. and Jhajj, P. (2010). Recognition of isolated handwritten characters in gurmukhi script. *International Journal of Computer Applications*, 4(8):9–17.

Siddharth, K. S., Dhir, R., and Rani, R. (2011). Handwritten gurmukhi numeral recognition using different feature sets. *International Journal of Computer Applications*, 28(2):20–24. Full text available.

Smith, E. H. B. (2010). An analysis of binarization ground truthing. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pages 27–34. ACM.

Smith, E. H. B. and An, C. (2012). Effect of" ground truth" on image binarization. In *Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on*, pages 250–254. IEEE.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 284 of 404

Trier, Ø. D., Jain, A. K., and Taxt, T. (1996). Feature extraction methods for character recognition-a survey. *Pattern Recognition*, 29(4):641 – 662.

Wahl, F. M., Wong, K. Y., and Casey, R. G. (1982). Block segmentation and text extraction in mixed text/image documents. *Computer graphics and image processing*, 20(4):375–390.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 285 of 404

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

In: Handwriting: Recognition, Development and Analysis ISBN: 978-1-53611-937-4
Editors: Byron L. D. Bezerra et al. © 2017 Nova Science Publishers, Inc.

Chapter 10

USING SPEECH AND HANDWRITING IN AN INTERACTIVE APPROACH FOR TRANSCRIBING HISTORICAL DOCUMENTS

Emilio Granell*[*], *Verónica Romero and Carlos-D. Martínez-Hinarejos
PRHLT Research Center, Universitat Politècnica de València,
Valencia, Spain

1. Introduction

Transcription of handwritten documents has become an interesting research topic in the last years. In particular, transcription of historical documents is interesting for preserving and providing access to data on cultural heritage (Fischer et al., 2009). Since accessibility to the contents of the documents is very limited without a proper transcription, this activity is needed to provide indexing, consulting and querying facilities on the contents of the documents.

The difficulties that historical manuscripts present make necessary the use of experts, called paleographers, that employ their knowledge on ancient script and vocabulary for obtaining an accurate transcription. In any case, this manual transcription is both slow and expensive. In order to make the process more efficient an interesting option is automatic transcription, which can employ the Handwritten Text Recognition (HTR) technology to obtain a transcription of the document. However, current state-of-the-art HTR technology does not guarantee an accurate enough transcription for the subsequent processes on the obtained data (Fischer et al., 2009; Serrano et al., 2010a), and paleographer intervention is required.

In order to alleviate the paleographer task on obtaining an accurate transcription from an initial HTR transcription, interactive assistive approaches have been introduced recently (Serrano et al., 2010b; Romero et al., 2012; Toselli et al., 2011; Llorens et al., 2009). In these approaches, the user and the system work together to obtain the perfect transcription; the system uses the text image, the automatic transcript provided by the HTR system

[*]E-mail address: egranell@dsic.upv.es (Corresponding author)

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 287 of 404

and some feedback given by the user to provide a new, hopefully, better hypothesis.

Apart from that, additional data sources could be provided to improve the initial transcription. For example, paleographers may employ speech dictation of the contents of the image to be transcribed. Dictation can be processed with Automatic Speech Recognition (ASR) techniques (Jelinek, 1998) and the recognition can be combined with HTR recognition results to obtain a more accurate transcription. This possibility was explored in (Granell and Martínez-Hinarejos, 2015a), using Confusion Networks (CN) combination; CN combination was previously studied for unimodal and multimodal signal integration with good results (Ishimaru et al., 2011; Granell and Martínez-Hinarejos, 2015a; Granell and Martínez-Hinarejos, 2015b). However, combination effects on interactive systems were not tested.

Additionally, in interactive systems, the user must provide feedback several times to the system, independently of the initial transcription given by the available data sources. Although the number of interactions may change depending on these initial sources, making the interaction process comfortable to the user is crucial to the success of an interactive system. Since paleographers usually employ touchscreen tablets for their task, using touchscreen pen strokes to provide feedback appears as an appropriate interactive option. These ideas have been previously explored in (Romero et al., 2012; Martín-Albo et al., 2013) in the context of a computer assisted transcription system called CATTI. However, this previous work employs a suboptimal two-phases process in each interaction step.

The work described in this chapter explores the effect of the combination of text images and speech signal as a new source for the interactive system, and the use of *on-line* text feedback that is integrated into each interaction in a single step by using CN combination. This feedback modality will be applied to the result of unimodal (text image or speech dictation) or multimodal recognition (combination of both text image and speech dictation recognition). The main hypothesis is that using more ergonomic multimodal interfaces should provide a more comfortable human-machine interaction, at the expense of employing a less deterministic feedback than when using not so ergonomic peripherals (e.g., keyboard and mouse). Thus, additional interaction steps may be necessary to correct possible errors produced when combining the current hypothesis and the feedback and their impact in productivity must be measured, specially for the multimodal source.

In summary, this chapter presents the use of combination of text images and speech signal as a new source to improve the initial hypothesis offered to the user of the interactive system, and the use of *on-line* text as a correction feedback, integrating it into the current transcription hypothesis. Results show how *on-line* HTR hypotheses can correct several errors on the initial hypotheses of the multimodal recognition process, providing a global reduction of the user effort, and thus allowing to speed up the transcriber task.

Section “Multimodal Interactive Transcription of Handwritten Text Images” presents the CATTI framework and the multimodal version of this approach. Section “Multimodal Combination in CATTI” explains the Confusion Network combination. Section “Natural Language Recognition Systems” gives an overview of the *off-line* HTR system, the ASR system, and the *on-line* HTR feedback subsystem. Section “Experimental Framework” details the experimental framework (data, conditions, and assessment measures); Section “Experimental Results” shows the results; and Section “Conclusions and Future Work” offers the final conclusions and future work lines.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 288 of 404

2. Multimodal Interactive Transcription of Handwritten Text Images

As previously commented, transcription of historical documents has become an interesting research topic in the last years. However, state-of-the-art handwritten text recognition systems (HTR) can not suppress the need of human work when high-quality transcriptions are needed. HTR systems can achieve fairly high accuracy for restricted applications with rather limited vocabulary (reading of postal addresses or bank checks) and/or form-constrained handwriting. However, in the case of historical handwritten documents, the current HTR technology typically only achieves results which do not meet the quality requirements of practical applications. Therefore, once the full recognition process of one document has finished, heavy human expert revision is required to really produce a transcription of standard quality. Such a *post-editing* solution is rather inefficient and uncomfortable for the human corrector.

A way of taking advantage of the HTR system is to combine it with the knowledge of a human transcriber, constituting the so-called "Computer Assisted Transcription of Text Images" (CATTI) scenario (Romero et al., 2012). In this framework, the automatic HTR system and the human transcriber cooperate interactively to obtain the perfect transcript of the text images. At each interaction step, the system uses the text image and a previously validated part (prefix) of its transcription to propose an improved output. Then, the user finds and corrects the next system error, thereby providing a longer prefix which the system uses to suggest a new, hopefully, better continuation.

Speech dictation of the handwritten text can be used as an additional or an alternative information source in the CATTI process. Taking into account both the handwritten text image and the speech signal, the system can, hopefully, propose a better transcription hypothesis in each interaction step. This way, many user corrections are avoided. Finally, in order to make the interaction more comfortable to the user, the feedback provided in each interaction step can be quite naturally provided by means of *on-line* text or pen strokes exactly registered over the text produced by the system.

In this section, we review the classical HTR and ASR framework and formalise the multimodal CATTI scenario where both sources, text and speech, help each other to improve the system accuracy. Finally, the multimodal approach where the feedback is provided by means of *on-line* text is also introduced.

2.1. HTR and ASR Framework

The traditional HTR and ASR recognition problems can be formulated in a very similar way. The problem is finding the most likely word sequence, $\hat{w}$, for a given handwritten sentence image or a speech signal represented by a feature vector sequence $x = (x_1, x_2, \ldots, x_{|x|})$ (Toselli et al., 2004), that is:

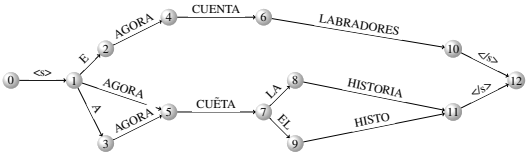
$$\hat{w} = \arg\max_{w \in W} \mathrm{P}(w \mid x) = \arg\max_{w \in W} \frac{\mathrm{P}(x \mid w)\mathrm{P}(w)}{\mathrm{P}(x)} = \arg\max_{w \in W} \mathrm{P}(x \mid w)\mathrm{P}(w) \qquad (1)$$

where W denotes the set of all permissible sentences, $\mathrm{P}(x)$ is the a priori probability of observing x, $\mathrm{P}(w)$ is the probability of $w = (w_1, w_2, \ldots, w_{|w|})$ approximated by the language

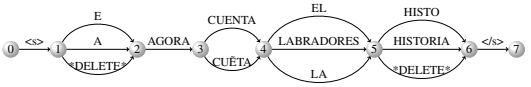
Cerence Operating Company, Ex. 2081
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 289 of 404

<s>	E	AGORA	CUENTA		LABRADORES	</s>
<s>		AGORA	CUẼTA	LA	HISTORIA	</s>
<s>		AGORA	CUẼTA	EL	HISTO	</s>
<s>	A	AGORA	CUẼTA	LA	HISTORIA	</s>
<s>	A	AGORA	CUẼTA	EL	HISTO	</s>

(a) N-best list.



(b) Word Graph.



(c) Confusion Network.

Figure 1. Recognition output formats: n-best list, Word Graph and Confusion Network.

model (usually modelled by a *n*-gram word language model), and $P(x \mid w)$ is the probability of observing x by assuming that w is the underlying word sequence for x, evaluated by the optical or acoustical models for HTR and ASR respectively (typically it is approximated by concatenated Hidden Markov Models - HMM).

The search (or decoding of) $\hat{w}$ is carried out by the Viterbi algorithm (Jelinek, 1998). From this dynamic-programming decoding process we can obtain not only a single best hypothesis, but also a huge set of best hypotheses. These solutions can be presented in the form of n-best list or compactly represented into Word Graphs (WG) or Confusion Networks (CN) (Jurafsky and Martín, 2009). A WG is a directed, acyclic and weighted graph that represents a huge set of hypotheses in a very efficient way. The nodes in a WG correspond to discrete time points for ASR and horizontal positions for HTR. The edges are labelled with words and weighted with the likelihood that the word appears in the signal delimited between the starting and ending nodes of the edge. The scores are derived from the HMM and *n*-gram probabilities computed during the decoding process. On the other hand, a CN is also a directed, acyclic and weighted graph that shows at each point which word hypotheses are competing or confusable. Each hypothesis goes through all the nodes. The words and their probabilities are stored in the edges, and the total probability of the words contained in a subnetwork (SN, all edges between two consecutive nodes), sum up to 1. Figure 1 provides an example of a n-best list, a WG representing these n-best hypotheses, and an equivalent CN.

2.2. CATTI Formal Framework

As previously explained, in the CATTI framework the user is directly involved in the transcription process, since he/she is responsible for validating and/or correcting the system hypothesis during the transcription process. The system takes into account the handwritten text image and the feedback of the user in order to improve these proposed hypotheses. The more information the system has about what is written in the handwritten text line image, the better the proposed hypotheses are, and therefore, fewer user interactions are needed to obtain the perfect transcript. In this work, in addition to the handwritten text line image, we study how the CATTI system can take advantage of the speech dictation of the text that the

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 290 of 404

images contain.

The process starts when the system proposes a full transcription $\hat{s}$ of the handwritten text line image taking into account also the speech dictation. Then, the user reads this transcription until finding a mistake and makes a mouse action (MA) m, or equivalent pointer-positioning keystrokes, to position the cursor at this point. By doing so, the user is already providing some very useful information to the system: he is validating a prefix p of the transcription (which is error-free) and, in addition, he is signalling that the following word e located after the cursor is incorrect. Hence, the system can already take advantage of this fact and directly propose a new suitable suffix (i.e., a new $\hat{s}$) in which the first word is different from the first wrong word of the previous suffix. This way, many explicit user corrections are avoided (Romero et al., 2008). If the new suffix $\hat{s}$ corrects the erroneous word, a new cycle starts. However, if the new suffix has an error in the same position than the previous one, the user can enter a word v to correct the erroneous one. This action produces a new prefix p (the previously validated prefix followed by the new word). Then, the system takes into account the new prefix to suggest a new suffix and a new cycle starts. This process is repeated until a correct transcription is accepted by the user.

In Figure 2 we can see an example of the CATTI process. In this example, without interaction, a user should have to correct about three errors from the originally recognised hypothesis ("abadia", "segun" and "el"). Using CATTI only one explicit user-correction is necessary to get the final error-free transcription: the interaction 1 only needs a MA, but in the interaction 2 a single mouse action does not succeed and the correct word needs to be typed.

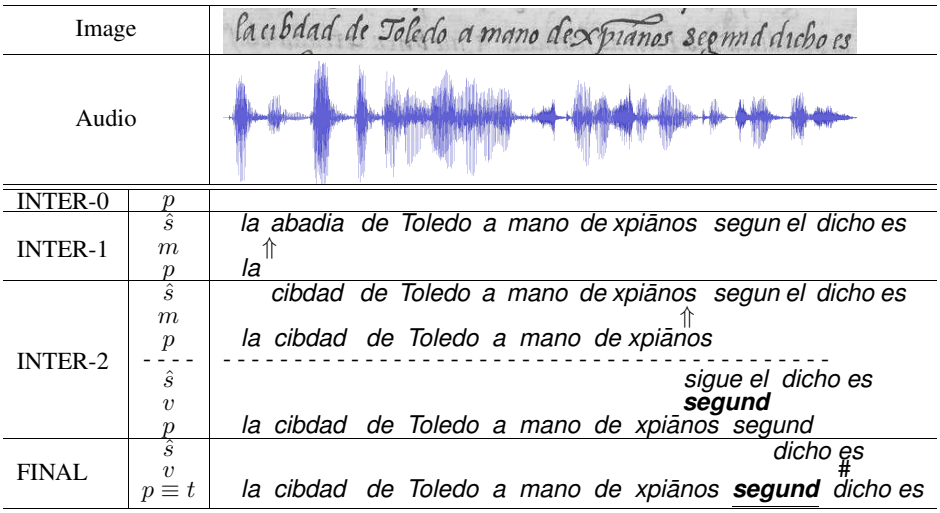


Figure 2. Example of CATTI operation using mouse-actions (MA).

Based on the Figure 2, the system starts with an initially recognised hypothesis $\hat{s}$ from any of the modalities or a combination of both, the user validates its longest well-recognised prefix p by making a MA m, and the system emits a new recognised hypothesis $\hat{s}$. As the new hypothesis corrects the erroneous word, a new cycle starts. Now, the user validates the

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 291 of 404

new longest prefix p which is error-free by making another MA m. The system provides a new suffix $\hat{s}$ taking into account this information. As the new suffix does not correct the mistake the user types the correct word v, generating a new validated prefix p. Taking into account the new prefix, the system suggests a new hypothesis $\hat{s}$. As the new hypothesis corrects the erroneous word, a new cycle starts. This process is repeated until the final error-free transcription t is obtained. The underlined boldface word in the final transcription is the only one which was corrected by the user. Note that in the second interaction (INTER-2) the correct word must be typed, given that the erroneous word was not corrected by the performed MA. However, in the first interaction (INTER-1) only a MA is needed to correct the erroneous word.

Formally, in the traditional CATTI framework (Romero et al., 2012), the system uses a given feature sequence, x_{htr}, representing a handwritten text line image and a user validated prefix p of the transcription. In this work, in addition to x_{htr}, we study how a sequence of feature vectors x_{asr} representing the speech dictation of the handwritten text line image affects the system performance. Therefore, the CATTI system should try to complete the validated prefix by searching for a most likely suffix $\hat{s}$ taking into account both sequences of feature vectors:

$$\hat{s} = \arg\max_{s} \mathrm{P}(s \mid x_{htr}, x_{asr}, p) \tag{2}$$

Making the naive assumption that x_{htr} does not depend on x_{asr}, and applying the Bayes' rule, we can rewrite the previous equation as:

$$\hat{s} = \arg\max_{s} \mathrm{P}(x_{htr} \mid p, s) \cdot \mathrm{P}(x_{asr} \mid p, s) \cdot \mathrm{P}(s \mid p) \tag{3}$$

where the concatenation of p and s is w in Equation 1. As in conventional HTR and ASR, $\mathrm{P}(x_{htr} \mid p, s)$ and $\mathrm{P}(x_{asr} \mid p, s)$ can be approximated by HMMs and $\mathrm{P}(s \mid p)$ by a n-gram model conditioned by p. Therefore, the search must be performed over all possible suffixes of p (Romero et al., 2012).

This suffix search can be efficiently carried out by using Word Graphs (WG) (Romero et al., 2012) or Confusion Networks (CN) (Granell et al., 2016). In each interaction step, the decoder parses the validated prefix p over the WG or CN and then continues searching for a suffix which maximises the posterior probability according to Equation (3). This process is repeated until a complete and correct transcription of the input text line image is obtained.

2.3. Multimodal Feedback

In the CATTI framework, users are repeatedly interacting with the system. Therefore, the quality and ergonomics of the interaction process are crucial for the success of the system. Traditional peripherals like keyboard and mouse can be used to unambiguously provide the feedback associated with the validation and correction of the successive system predictions. Nevertheless, using more ergonomic multimodal interfaces should result in an easier and more comfortable human-machine interaction, at the expense of a less deterministic feedback. It is important to note that the use of this more ergonomic user interaction will produce new errors coming from the decoding of the feedback signals. Here, we will focus on touchscreen communication, which is perhaps the most natural feedback modality for

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 292 of 404

multimodal CATTI (MM-CATTI). In this way, the user corrective feedback can be quite naturally provided by means of *on-line* text or pen strokes exactly registered over the text produced by the system.

In (Romero et al., 2012), the multimodal interaction process is formulated into two steps. In the first step a CATTI system solves the problem presented in Equation (3). In the second step, the user enters some pen-strokes, t, typically aimed at accepting or correcting parts of the suffix suggested by the system in the previous interaction step, $\hat{s}$, validating a prefix which is error free, p'. Then, an *on-line* HTR feedback subsystem is used to decode t into a word $\hat{d}$, taking into account $\hat{s}$ and p'. This scenario is considered here as a baseline.

In (Granell et al., 2016), a multimodal interaction process in one step was presented. In this way, both the CATTI source (it can be only *off-line* text, speech or both) and *on-line* handwritten text help each other to optimise the system accuracy.

Formally speaking, let x_{htr} be the input image, x_{asr} the speech dictation of the input image, and t the *on-line* touchscreen pen strokes that the user introduces to insert or substitute a word. Let p' be the user-validated prefix of the previously suggested transcription which is error-free and e the wrong word that the user tries to correct. Using this information, the system has to suggest a new suffix, s, as a continuation of the validated prefix p', conditioned by the *on-line* touchscreen strokes t and the erroneous word e. Therefore, the problem is to find $\hat{s}$ given x_{htr}, x_{asr} and a feedback information composed of p', e and t. By further considering the decoding d of t as a hidden variable, we can write:

$$\begin{aligned}\hat{s} &= \arg\max_{s} \sum_{d} \mathrm{P}(s,d \mid x_{htr}, x_{asr}, p', t, e) \\ &\approx \arg\max_{s} \sum_{d} \mathrm{P}(t \mid p', e, s, d, x_{htr}, x_{asr}) \cdot \mathrm{P}(x_{htr} \mid p', e, s, d) \cdot \mathrm{P}(x_{asr} \mid p', e, s, d) \\ &\quad \cdot \mathrm{P}(s \mid p', e, d) \cdot \mathrm{P}(d \mid p', e) \end{aligned} \tag{4}$$

We can now make the reasonable assumption that t only depends on d and, that x_{htr}, x_{asr} and s do not depend on e and, approximating the sum over all the possible decodings d of t by the dominating term, Equation (4) can be rewritten as:

$$\hat{s} \approx \arg\max_{s} \max_{d} \mathrm{P}(x_{htr} \mid p', d, s) \cdot \mathrm{P}(x_{asr} \mid p', d, s) \cdot \mathrm{P}(s \mid p', d) \cdot \mathrm{P}(t \mid d) \cdot \mathrm{P}(d \mid p', e) \tag{5}$$

The first three terms of Equation (5) are very similar to Equation (3), being p the concatenation of p' and d. The main difference is that now d is unknown. On the other hand, the last two terms correspond to the HTR decoding of the *on-line* feedback, conditioned by the previously validated prefix p' and the erroneous word e. As in conventional CATTI, the probabilities $\mathrm{P}(x_{htr} \mid p', d, s)$, $\mathrm{P}(x_{asr} \mid p', d, s)$ and $\mathrm{P}(t \mid d)$ are modelled by optical, acoustical and kinematical HMMs, whereas, $\mathrm{P}(s \mid p', d)$ and $\mathrm{P}(d \mid p', e)$ are modelled by using conditioned n-grams.

In order to cope with the erroneous word e that follows the validated prefix, and given that this word only affects to the decoding of t, $\mathrm{P}(d \mid p', e)$ can be formulated as follows:

$$\mathrm{P}(d \mid p', e) = \frac{\overline{\delta}(d, e) \cdot \mathrm{P}(d \mid p')}{1 - \mathrm{P}(e \mid p')} \tag{6}$$

where $\overline{\delta}(i, j)$ is 0 when $i = j$ and 1 otherwise.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 293 of 404

As in conventional CATTI, this decoding can be implemented using CN. In each interaction step, the validated prefix p' is parsed over the CN obtained from the combination of the *off-line* original image x_{htr} and/or the speech dictation x_{asr}. This parsing procedure will end defining a node q of the CN whose associated word sequence is p'. Then, a CN is obtained from the *on-line* feedback recogniser. Assuming that the user corrects only one word in each interaction, this *on-line* CN is composed of a list of words that corresponds with the different decodings of t. This *on-line* CN is combined with the *off-line* CN after the node q. Then, the system continues searching for the most probably suffix, according to Equation (5), using this new combined CN.

3. Multimodal Combination in CATTI

Since *off-line* HTR and ASR systems share most part of the recognition process (see Section "Natural Language Recognition Systems"), the possibility of using the results of both systems in CATTI arises immediately. In this way, the CATTI system would take advantage from two different data sources. Moreover, the integration of the more ergonomic *on-line* HTR feedback provides CATTI with an additional source of information for correcting specific errors.

In this work, Confusion Networks (CN) are used in the search process of CATTI because they reduce the complexity of Word Graphs without losing important information (Xue and Zhao, 2005) and allow to carry out a multimodal combination based on the Bayes theorem and assuming a strong independence between the three modalities with an insignificant runtime (Granell and Martínez-Hinarejos, 2015a; Granell and Martínez-Hinarejos, 2015b).

3.1. Multimodal Hypotheses Combination

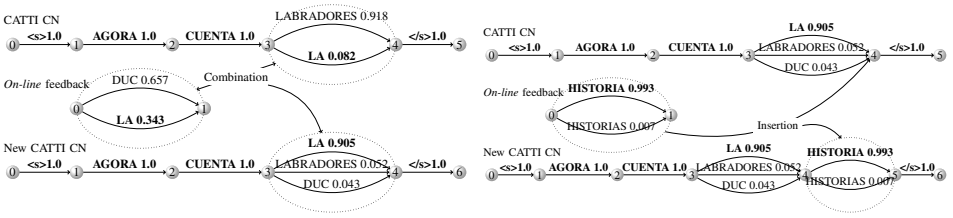
The bimodal CN combination method (Granell and Martínez-Hinarejos, 2015a) used to combine the two initial sources of information in CATTI works as follows, starting from the decoding outputs of the *off-line* HTR and ASR recognisers formatted in CN:

1. A search for anchor subnetworks is performed in order to align the subnetworks of both CN. The algorithm searches coincidences in unigrams, bigrams and skip-bigrams in both directions (from left to right and vice versa) simultaneously, taking only as anchor subnetworks those subnetworks where both searches coincide, according to a gram matching value of the words in the involved subnetworks. This gram matching is assessed by using the quadratic mean of the Character Error Rate (CER) and the Phoneme Error Rate (PER) between those words.

$$\mathrm{E}(w_{HTR}, w_{ASR}) = \sqrt{\frac{\mathrm{CER}(w_{HTR}, w_{ASR})^2 + \mathrm{PER}(w_{HTR}, w_{ASR})^2}{2}} \tag{7}$$

   where w_{HTR} and w_{ASR} represent the words of the subnetworks of both modalities. CER and PER are the Levensthein distance between those words, CER at character level, and PER at phoneme level by using the phonetic transcriptions of the recognized words, and E represents the gram matching error.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 294 of 404



(a) Subnetwork combination example.

(b) Subnetwork insertion example.

Figure 3. MM-CATTI editing actions. Ref.: <s>AGORA CUENTA LA HISTORIA </s>.

2. The new CN is composed by using three editing actions: combination, insertion and deletion of subnetworks:

   Combination: Given two subnetworks, SN_{HTR} and SN_{ASR}, the word posterior probabilities of the combined CATTI subnetwork SN_{CATTI} are obtained by applying a normalisation on the logarithmic interpolation of the smoothed word posterior probabilities of both SN (SN_{HTR} and SN_{ASR}):

   $$\Pr(w \mid SN_{CATTI}) = \Pr_s(w \mid SN_{HTR})^{\alpha} \Pr_s(w \mid SN_{ASR})^{1-\alpha} \tag{8}$$

   where the weight factor α allows to balance the reliability between modalities, and the smoothing of the word posterior probabilities is calculated according to the following equation which is based on Laplacian smoothing:

   $$\Pr_s(w \mid SN) = \frac{\Pr(w \mid SN) + \Theta}{1 + n\Theta} \tag{9}$$

   where Θ is a defined granularity that represents the minimum probability for a word and n is the number of different words in the final CATTI SN.

   Insertion and deletion: The same process is performed in both actions: the subnetwork to insert or to delete is combined with a subnetwork with an only *DELETE* arc with probability 1.0.

3.2. Multimodal Hypotheses Correction

As the *on-line* HTR feedback is limited to one word, the *on-line* CN obtained is composed by only two nodes, like a subnetwork. This *on-line* CN is combined or inserted into the CATTI CN at the point that the previous parsing of the user validated prefix has defined. Therefore, two different editing operations can be carried out to generate the new CATTI CN: subnetwork combination and subnetwork insertion:

Combination: The same combination process explained for the multimodal hypotheses combination is performed in this case.

In the example (Figure 10.3a), the marked subnetwork of the CATTI CN (subnetwork between the nodes 3 and 4) is selected for combination. In this case, the correct word

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 295 of 404

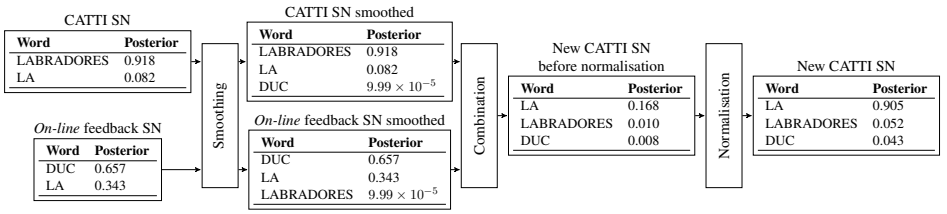


Figure 4. Detailed subnetwork combination example using $\alpha = 0.5$ and $\Theta = 10^{-4}$. The Smoothing block represents the use of Equation (9), and the Combination block the use of Equation (8).

(*LA*) is not the most probable word, either in the *on-line* subnetwork. However, it becomes the most probable word when combining both subnetworks with $\alpha = 0.5$ and $\Theta = 10^{-4}$, as can be seen in the appointed subnetwork of the new CN (Figure 4 shows this combination process in detail).

Insertion: The subnetwork insertion allows adding a word into the CATTI CN on a particular position. This position is determined by the parsing of the validated prefix p' that precedes the *on-line* word inserted by the user in the CATTI interaction.

As an example, the *on-line* SN (see Figure 3) is inserted just after the 4$^{\text{th}}$ node of the CATTI CN.

4. Natural Language Recognition Systems

A similar conceptual architecture composed of three modules (*preprocessing*, *feature extraction* and *recognition*) is adopted for the three different recognition systems, *off-line* HTR, ASR, and *on-line* HTR systems.

On the one hand, *off-line* HTR preprocessing is aimed at correcting image degradations and geometry distortions, while *on-line* HTR preprocessing involves only two simple steps: repeated points elimination and noise reduction. Feature extraction in the *off-line* HTR case transforms a preprocessed text line image into a sequence of 60-dimensional feature vectors, whereas for *on-line* HTR a touchscreen coordinates sequence is transformed into a new speed- and size-normalised temporal sequence of 6-dimensional real-valued feature vectors (Toselli et al., 2007).

On the other hand, ASR preprocessing is inspired by the human auditory physiology and perception. The extraction of Mel-Frequency Cepstral Coefficient (MFCC) speech features, transforms an audio sequence into a sequence of 39-dimensional feature vectors (Rabiner and Juang, 1993).

The recognition process is the same for the three systems. Characters and phonemes are modelled by continuous density left-to-right HMMs. A gaussian mixture serves as a probabilistic law to model the emission of feature vectors of each HMM state. On the other hand, each lexical word is modelled by a stochastic finite-state automaton, which represents all possible concatenations of individual characters or phonemes to compose the word. Finally, text sentences are modelled using word 2-grams with Kneser-Ney back-off

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 296 of 404

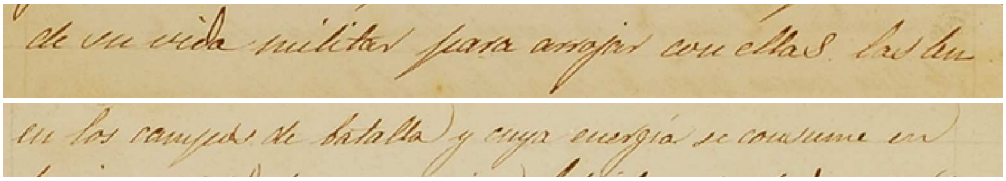


Figure 5. Sample lines for *Cristo Salvador*.

smoothing (Kneser and Ney, 1995), estimated directly from the training transcriptions of the text line images. The decoding is optimally carried out by the Viterbi algorithm. This dynamic-programming decoding process can return not only a single best solution, but also a huge set of best solutions compactly represented into a WG. Then this WG is transformed into a CN. For a more detailed description of these recognition systems see (Toselli et al., 2004) for *off-line* HTR, (Rabiner and Juang, 1993) for ASR, and (Toselli et al., 2007) for *on-line* HTR.

5. Experimental Framework

5.1. Datasets

5.1.1. *Off-line* Handwritten Text: *Cristo Salvador*

The *Cristo Salvador* corpus was employed previously in different works (Alabau et al., 2011, 2014; Granell and Martínez-Hinarejos, 2015a) related to multimodal combination. This corpus is a handwritten book of the XIX century provided by *Biblioteca Valenciana Digital* (BiValDi). It is a single writer book with different image features that cause some problems, such as smear, background variations, differences in bright, and bleed-through (ink that trespasses to the other surface of the sheet). It is composed of 53 pages that were automatically divided into lines (such as shown in Figure 5).

This corpus presents a total number of 1,172 lines, with a vocabulary of 3,287 different words. For training the optical models for *off-line* HTR, a partition with the first 32 pages (675 lines) was used. Test data for *off-line* HTR is composed of the lines of page 41 (24 lines, 222 words), that was selected for being, according to preliminary error recognition results, a representative page of the whole test set (the remaining 21 pages, 497 lines).

5.1.2. Speech: *Albayzin* and *Cristo Salvador*

With respect to the acoustical models for ASR, a partition of the *Albayzin* Spanish database (Moreno et al., 1993) was used for training them. This corpus consists of a set of three sub-corpus recorded by 304 speakers using a sampling rate of 16 kHz and a 16-bit quantisation. The training partition used in this work includes 4800 phonetically balanced utterances. Specifically, 200 utterances were read by four speakers and 25 utterances were read by 160 speakers, with a total length of about 4 hours.

Test data for ASR was the product of the acquisition of the dictation of the contents of the lines of page 41 of *Cristo Salvador* by five different native Spanish speakers (i.e. a total

Cerence Operating Company, Ex. 2041
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 297 of 404



Figure 6. Examples of the word “HISTORIA” generated by using characters from the three selected UNIPEN test writers (BH, BR, BS).

of 120 utterances, with an average length of about 4 and a half seconds), using a sample rate of 16 KHz and an encoding of 16 bits (to match the conditions of *Albayzin* data).

5.1.3. *On-line* Handwritten Text: UNIPEN

The production of touchscreen feedback data has been simulated using the UNIPEN `Train-R01/V07` dataset[1]. It comes organised into several categories such as lower and upper-case letters, digits, symbols, isolated words and full sentences. Unfortunately, the UNIPEN isolated words category does not contain the required word instances to be handwritten by the user in the MM-CATTI interaction process with *Cristo Salvador* text images. Therefore, the process for generating synthetic samples used in (Romero et al., 2012) was followed here. The samples were generated by concatenating random character instances from three UNIPEN categories: `1a` (digits), `1c` (lowercase letters) and `1d` (symbols).

To increase realism, the generation of each of these test words was carried out employing characters belonging to the same writer. Three different writers were randomly chosen, taking care that sufficient samples of all the characters needed for the generation of the required word instances were available from each writer. Each character needed to generate a given word was plainly aligned along a common word baseline, except if it had a descender, in which case the character baseline was raised 1/3 of its height. The horizontal separation between characters was randomly selected from *one* to *three* trajectory points. The selected writers are identified by their name initials as BS, BH and BR. Figure 6 presents three examples of the word “HISTORIA” for the three different writers generated in this way.

Training data was produced in a similar way using 17 different UNIPEN writers. For each of these writers, a sample of each of the 42 symbols and digits needed was randomly selected and one sample of each of the $1,000$ most frequent Spanish and English words was generated, resulting in $34,714$ training tokens (714 isolated characters plus $34,000$ generated words). To generate these tokens, $186,881$ UNIPEN character instances were used, using as many repetitions as required out of the $17,177$ unique character samples available. Table 1 summarises the amount of UNIPEN training and test data used in our experiments.

It should be mentioned here that, even though the *on-line* HTR kinematical models were trained from artificially built words, the accuracy in real operation with real users performed in (Romero et al., 2012) is observed to be similar to that shown in the laboratory results reported with synthetic samples.

[1]For a detailed description of this dataset, see http://www.unipen.org.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 298 of 404

Table 1. Basic statistics of the UNIPEN training and test data used in the experiments.

Number of different:	Train	Test	Lexicon
writers	17	3	-
digits (1a)	1,301	234	10
letters (1c)	12,298	2,771	26
symbols (1d)	3,578	3,317	32
total characters	17,177	6,322	68

5.2. System Setup

Optical, acoustical, and kinematical models were trained by using HTK (Young et al., 2006). In the first place, each *off-line* character is modelled on the optical model by a continuous density left-to-right HMM with 6 states and 32 gaussians per state. Secondly, phonemes are modelled on the acoustical model as a left-to-right HMM with 3 states and 64 gaussians per state. And finally, a variable number of states for the different *on-line* characters was used with 16 gaussians per state on the kinematical model.

The Language Model (LM) was estimated as a *2*-gram with Kneser-Ney back-off smoothing (Kneser and Ney, 1995) directly from the transcriptions of the 32 pages included in the *off-line* HTR training set. This LM was interpolated with the whole lexicon in order to avoid out-of-vocabulary words, and it presents a perplexity of 742.8 for the test data.

The three recognition systems were implemented by using the iATROS recognizer (Luján-Mares et al., 2008), and the SRILM toolkit (Stolcke, 2002) was used for all processes on language models, and for obtaining CN from the WG of the decoding outputs.

In order to optimize the experiments results, the values of the main variables were tuned. In the CATTI and MM-CATTI experiments, the limit of mouse actions was set to 5, and for the multimodal combination, a weight factor of $\alpha = 0.5$ and a granularity factor of $\Theta = 10^{-4}$ were used.

5.3. Evaluation Metrics

Different evaluation measures have been adopted. On the one hand, the quality of the transcription without any system-user interaction is given by the well known word error rate (WER), which is a good estimation of the user post-edition effort. It is defined as the minimum number of words to be substituted, deleted or inserted to convert the hypothesis into the reference, divided by the total number of reference words. In addition, the oracle WER represents the best WER that can be obtained from the WG.

On the other hand, the CATTI performance is given by the word stroke ratio (WSR), which can be also computed using the reference transcription. After each CATTI hypothesis, the longest common prefix between the hypothesis and the reference is obtained and the first mismatching word from the hypothesis is replaced by the corresponding reference word. This process is iterated until a full match is achieved. Therefore, the WSR can be defined as the number of user interactions that are necessary to produce correct transcriptions using the CATTI system, divided by the total number of reference words. This definition

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 299 of 404

makes WER and WSR comparable. The relative difference between them gives us a good estimation of the reduction in human effort (EFR) that can be achieved by using CATTI with respect to using a conventional HTR system followed by human post-editing.

Moreover, since *on-line* feedback is used only for single-word corrections, the conventional classification error rate (ER) was used to assess its recognition quality.

For each measure, confidence intervals of 95% were calculated by using the bootstrapping method with 10,000 repetitions (Bisani and Ney, 2004).

6. Experimental Results

Several experiments were performed to assess our multimodal proposal for improving the assistive transcription system presented in Section "Multimodal Interactive Transcription of Handwritten Text Images". Multimodal combination allows to enrich the CATTI hypotheses from different sources of information (*off-line* HTR and ASR). Moreover, the multimodal operation with the *on-line* HTR feedback offers ergonomics and increased usability at the expense of the system having to deal with non-deterministic feedback signals. Therefore, the main concern here is how the performance in CATTI and MM-CATTI can be boosted by the multimodal combination of different decoding systems. Finally, we assess which degree of synergy can be expected by taking into account both interaction and multimodality.

We started obtaining the non interactive post-edition baseline, for the *off-line* HTR, for the ASR, and for the multimodal combination of both unimodal recognition systems. Then, the CATTI and the MM-CATTI approaches were applied to the three input possibilities, two unimodal (*off-line* HTR and ASR), and one multimodal (*off-line* HTR combined with ASR) formatted as CN.

6.1. Post-Edition Baseline Results

Table 2. Post-Edition Experimental Results.

Measure	*Off-line* HTR	ASR	Multimodal
WER	$32.9\% \pm 6.4$	$43.7\% \pm 3.3$	$29.3\% \pm 2.5$
Oracle WER	$27.5\% \pm 6.4$	$27.4\% \pm 2.2$	$13.4\% \pm 2.1$
Time per sample (ms)	204642.2	30144.5	204957.7

In Table 2 the baseline results are presented. This table shows the WER and the oracle WER obtained during the conventional, non-interactive experiments performed on the *Cristo Salvador* dataset. As can be observed in the post-edition results, the *off-line* HTR decoding output presents $32.9\% \pm 6.4$ of WER with an oracle WER value of $27.5\% \pm 6.4$. Regarding the ASR obtained results, speech recognition does not seems to be a good substitute for handwriting recognition in this task, although both modalities present similar oracle WER values. Given that these unimodal oracle WER values are not significantly better than the *off-line* HTR baseline WER value, a significant effort reduction produced by the CATTI and MM-CATTI systems can not be expected.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 300 of 404

However, the multimodal combination of both sources allows to reduce the WER value to 29.3% ± 2.5, which represents a relative improvement of 10.9% over the *off-line* HTR baseline, and 33.0% over the ASR baseline. One of the best advantages of the multimodal combination is that not only the 1-best is improved, but also the rest of hypotheses are improved. This fact can be observed through the oracle WER of this multimodal combination (13.4% ± 2.1), which is significantly reduced given the two unimodal sources. Given that the oracle WER represents the WER of the best hypothesis that can be achieved, a significant beneficial effect on interactive systems can be expected.

Concerning the time processing performance, as can be observed in Table 2, the *off-line* HTR system needed on average 204642.2 ms (approximately 3.4 minutes) to decode a text line image, while the ASR system only needed 30.1 seconds to decode every speech utterance. On the other hand, the combination of the decoding results of both systems took on average 315.5 ms per sample. Taking into account that both decoding processes can be performed in parallel, the average time required for obtaining the multimodal output corresponds to the processing time of the slowest modality (in this case the *off-line* HTR), plus the combination time, i.e. 204957.7 ms on average. The multimodal combination represents a negligible increase of 0.2% of extra time over the time required to decode the *off-line* HTR.

6.2. CATTI Results

Table 3. CATTI Experimental Results.

Measure	CATTI Input		
	Off-line	**ASR**	**Multimodal**
WSR	31.1% ± 6.0	31.6% ± 3.1	12.9% ± 2.1
EFR	5.5%	4.0%	60.8%

Table 3 presents the estimated interactive human effort (WSR) required for obtaining the perfect transcription using the interactive CATTI approach for the three different input possibilities.

As expected, the obtained WSR for the unimodal inputs represents a slight effort reduction (EFR) of around 5% with respect to the *off-line* HTR baseline. However, in the case of the multimodal input the WSR reaches 12.9% ± 2.1, which represents a significant effort reduction of 60.8% over the *off-line* HTR baseline (32.9% ± 6.4). Notice that, in the multimodal case, the obtained WSR value is a bit lower than the oracle WER value (13.4% ± 2.1); this is possible because the presented CATTI approach, by means of mouse actions, allows reducing the number of words explicitly corrected by the user. Therefore, in this case, the CATTI approach not only offers the best hypothesis contained in the multimodal lattices, but it improves the oracle WER value deleting several erroneous words of this hypothesis.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 301 of 404

Table 4. *On-line* HTR Feedback Results.

Measure	MM-CATTI Input		
	***Off-line* HTR**	**ASR**	**Multimodal**
On-line Words	57.7	70.4	23.9
On-line HTR ER	$6.9\% \pm 3.2$	$3.6\% \pm 0.9$	$8.7\% \pm 2.4$

6.3. MM-CATTI Results

In MM-CATTI, the *on-line* HTR feedback results (see Table 4) were obtained decoding only the words that the user must introduce during the MM-CATTI process (in average, 57.7 words when the input was the unimodal *off-line* HTR, 70.4 words when the input was the unimodal ASR, and 23.9 words when the input was multimodal).

The *on-line* HTR feedback presented moderated writer average decoding error rates (ER). As presented in Table 4, the more words are decoded, the better the decoding error rates. This is due to the fact that the MM-CATTI input with worst hypotheses (ASR) needs the *on-line* HTR feedback to correct easier words (so there is less error), while the input with better hypotheses (Multimodal) needs the *on-line* HTR feedback to correct more difficult words (biggest mistake in proportion, although the overall number of words to correct is lower).

Table 5. MM-CATTI Experimental Results.

Measure		MM-CATTI Input		
		***Off-line* HTR**	**ASR**	**Multimodal**
WSR	TS	$26.0\% \pm 5.5$	$31.6\% \pm 3.4$	$10.7\% \pm 1.8$
	KBD	$12.3\% \pm 5.0$	$8.6\% \pm 1.5$	$3.2\% \pm 1.0$
	Global	$38.3\% \pm 8.6$	$40.2\% \pm 4.4$	$13.9\% \pm 2.4$
EFR		-16.4%	-22.2%	57.8%

In Table 5 the MM-CATTI results are presented. In this case, the estimated interactive human effort (WSR) is decomposed into the percentage of words written with the *on-line* HTR feedback - TouchScreen (TS) - and the percentage of those words for which the correction with the *on-line* HTR feedback failed and the corrections had to be entered by means of the keyboard (KBD), i.e., in MM-CATTI the WSR is calculated under the assumption that the cost of keyboard-correcting an erroneous *on-line* feedback word is similar to another *on-line* HTR interaction. This is a pessimistic assumption since interaction through touchscreen is more ergonomic than through the keyboard.

The multimodal combination of the *on-line* feedback with the MM-CATTI hypotheses allowed to reduce significantly a number of words that required to be corrected by using the keyboard. Despite the fact that in the unimodal input experiments only 12.3% of words for *off-line* HTR, and 8.6% for ASR were corrected by using the keyboard, no EFR can be considered given the previous definition of WSR for MM-CATTI. However, with the multimodal input a 13.9% of WSR was obtained, which represents an EFR of 57.8% with respect to the *off-line* HTR baseline.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 302 of 404

According to these results, in MM-CATTI most of the user effort is concentrated in the more ergonomic and user preferred touchscreen feedback. However, the overall user effort in MM-CATTI is only moderately higher than that of CATTI when the input presents a low oracle WER value.

Conclusion and Future Work

In this chapter, we have presented the use of Confusion Networks combination for improving the interaction (by using *on-line* touch-screen handwritten pen strokes) in an interactive transcription system presented in previous works. The main advantage of the presented approach is that the multimodal combination allows to correct errors on the MM-CATTI hypothesis by using the information provided by the *on-line* handwritten text introduced by the user.

The obtained results show the benefits of using speech as an additional source of information for the transcription of historical manuscripts. Moreover, the use of the more ergonomic feedback (*on-line* HTR) modality comes at the cost of only a reasonably small number of additional interaction steps needed to correct the few feedback decoding errors.

Our future works aim at taking advantage of the real samples that are produced while the system is used for adapting the *on-line* recogniser to the user. Moreover, we propose for future work the use of sentences in the *off-line* handwritten corpus instead of lines, in order to make the dictation of the contents more natural.

Acknowledgments

Work partially supported by projects READ - 674943 (European Union's H2020), SmartWays - RTC-2014-1466-4 (MINECO), and CoMUN-HaT - TIN2015-70924-C2-1-R (MINECO/FEDER).

References

Alabau, V., Martínez-Hinarejos, C.-D., Romero, V., and Lagarda, A. L. (2014). An iterative multimodal framework for the transcription of handwritten historical documents. *Pattern Recognition Letters*, 35:195–203.

Alabau, V., Romero, V., Lagarda, A. L., and Martínez-Hinarejos, C.-D. (2011). A Multimodal Approach to Dictation of Handwritten Historical Documents. In *Proc. 12th Interspeech*, pages 2245–2248.

Bisani, M. and Ney, H. (2004). Bootstrap estimates for confidence intervals in ASR performance evaluation. In *Proc. ICASSP*, volume 1, pages 409–412.

Fischer, A., Wuthrich, M., Liwicki, M., Frinken, V., Bunke, H., Viehhauser, G., and Stolz, M. (2009). Automatic transcription of handwritten medieval documents. In *2009 15th International Conference on Virtual Systems and Multimedia*. Institute of Electrical and Electronics Engineers (IEEE), pages 137-142.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 303 of 404

Granell, E. and Martínez-Hinarejos, C.-D. (2015a). Combining Handwriting and Speech Recognition for Transcribing Historical Handwritten Documents. In *Proc. 13th ICDAR*, pages 126–130.

Granell, E. and Martínez-Hinarejos, C.-D. (2015b). Multimodal Output Combination for Transcribing Historical Handwritten Documents. In *Proc. 16th CAIP*, pages 246–260.

Granell, E., Romero, V., and Martinez-Hinarejos, C.-D. (2016). An Interactive Approach with *Off-line* and *On-line* Handwritten Text Recognition Combination for Transcribing Historical Documents. In *Proc. DAS*, pages 269–274.

Ishimaru, S., Nishizaki, H., and Sekiguchi, Y. (2011). Effect of Confusion Network Combination on Speech Recognition System for Editing. In *Proc. 3rd APSIPA ASC*, volume 4, pages 1–4.

Jelinek, F. (1998). *Statistical Methods for Speech Recognition*. MIT Press.

Jurafsky, D. and Martín, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall, 2nd edition.

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Proc. ICASSP*, volume 1, pages 181–184.

Llorens, D., Marzal, A., Prat, F., and Vilar, J. M. (2009). State: an assisted document transcription system. In *Proc. ICMI-MLMI*, pages 233–234.

Luján-Mares, M., Tamarit, V., Alabau, V., Martínez-Hinarejos, C.-D., Pastor, M., Sanchis, A., and Toselli, A. H. (2008). iATROS: A speech and handwritting recognition system. In *Proc. V JTH*, pages 75–78.

Martín-Albo, D., Romero, V., and Vidal, E. (2013). Interactive off-line handwritten text transcription using on-line handwritten text as feedback. In *Proc. ICDAR*, pages 1280–1284.

Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J. B., and Nadeu, C. (1993). Albayzin speech database: design of the phonetic corpus. In *Proc. EuroSpeech*, pages 175–178.

Rabiner, L. R. and Juang, B. H. (1993). *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, New Jersey, USA.

Romero, V., Toselli, A. H., Civera, J., and Vidal, E. (2008). Improvements in the Computer Assisted Transcription System of Handwritten Text Images. In *Proc. 8th PRIS*, pages 103–112.

Romero, V., Toselli, A. H., and Vidal, E. (2012). *Multimodal Interactive Handwritten Text Transcription*, volume 80 of *Machine Perception and Artificial Intelligence*. World Scientific.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 304 of 404

Serrano, N., Castro, F., and Juan, A. (2010a). The RODRIGO Database. In *Proc. 7th LREC*, pages 2709–2712.

Serrano, N., Sanchis, A., and Juan, A. (2010b). Balancing Error and Supervision Effort in Interactive-Predictive Handwritten Text Recognition. In *Proc. 15th IUI*, pages 373–376.

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *Proc. 3rd Interspeech*, pages 901–904.

Toselli, A. H., Juan, A., Keysers, D., González, J., Salvador, I., H. Ney, Vidal, E., and Casacuberta, F. (2004). Integrated Handwriting Recognition and Interpretation using Finite-State Models. *IJPRAI*, 18(4):519–539.

Toselli, A. H., Vidal, E., and Casacuberta, F., editors (2011). *Multimodal Interactive Pattern Recognition and Applications*. Springer, 1st edition.

Toselli, A. H., Pastor, M., and Vidal, E. (2007). On-Line Handwriting Recognition System for Tamil Handwritten Characters. In *Proc. 3rd IbPRIA*, volume 4477, pages 370–377.

Xue, J. and Zhao, Y. (2005). Improved Confusion Network Algorithm and Shortest Path Search from Word Lattice. In *Proc. 30th ICASSP*, volume 1, pages 853–856.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al. (2006). The HTK book. *Cambridge university engineering department*.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 305 of 404

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

In: Handwriting: Recognition, Development and Analysis ISBN: 978-1-53611-937-4
Editors: Byron L. D. Bezerra et al. © 2017 Nova Science Publishers, Inc.

Chapter 11

HANDWRITTEN KEYWORD SPOTTING THE QUERY BY EXAMPLE (QBE) CASE

Georgios Barlas, Konstantinos Zagoris and Ioannis Pratikakis*
Department of Electrical and Computer Engineering
Democritus University of Thrace, Xanthi, Greece

1. Introduction

The traditional approach in document indexing usually involves an Optical Character Recognition (OCR) step. Although OCR performs well in modern printed documents and documents of high quality printing, in the case of handwritten documents OCR, several factors affect the final performance like intense degradation, paper-positioning variations (skew, translations, etc.) and writing styles variety.

Handwritten keyword spotting has attracted the attention of the research community in the field of document image analysis and recognition since it appears to be a feasible solution for indexing and retrieval of handwritten documents in the case that OCR-based methods fail to deliver satisfactory results.

Handwritten keyword spotting (KWS) is the task of retrieving all instances of a given query word in handwritten document image collections without involving a traditional OCR step. There exist two basic variations for KWS approaches: (a) the Query by Example case (QbE) where the query is a word image and (b) the Query by String case (QbS) where, as the name implies, the query is a string. The study presented in this chapter will focus on the QbE approach.

For a better understanding, QbE methods will be presented taking into account two different perspectives which relate to the use of segmentation and learning. The segmentation-based methods are divided into 2 subcategories based upon the segmented entity which could be either the word image or the text line. They are strongly dependent on the segmentation step so that to compare different methods regardless of segmentation errors, many researchers do not implement a segmentation method but they use datasets where the segments are given.

*E-mail address: ipratika@ee.duth.gr (Corresponding author).
Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

In the case of segmentation-free methods, the whole image is tested against similarities between the query image and the patches of the document image without segmenting it at any level. The methods of this class, on the one hand bypass the step of segmentation but on the other hand they cannot avoid searching for the words in parts of the image that may not contain text. Therefore, segmentation-free methods avoid failures due to bad segmentation but the running time increases considerably. It is worth-mentioning that the methods of this class are not the trend.

Training-based methods are those that require training data at a particular stage of the process. A common problem in these methods is the availability of training data. Furthermore, an extra weakness is that to apply such a method to a new word, usually ground truthing work is required to obtain training data, which is quite time consuming and often it has to be done totally manual.

Training - free are methods that as the name implies do not include any training stage in the operational KWS pipeline. The training - free methods can be applied directly to new word although, they usually require a particular configuration to be effective in the corresponding text.

This chapter is structured as follows: Section “Segmentation-based Context” will present the KWS methodologies that operate in a segmentation-based context wherein methods based on training and methods that are independent of any training involvement will be detailed. Both variations will be separately reviewed depending on the type of segmentation which is used. In Section “Segmentation - Free Context”, methodologies that do account for a segmentation will be discussed with a particular focus on the use or not of training. Section “Experimental Datasets and Evaluation Metrics” deals with an overview of the current efforts for performance evaluation and a brief description of datasets that were used in QbE KWS, while the Section “Concluding Remarks” is dedicated to a fruitful discussion which aims to identify the current trends of the QbE KWS.

2. Segmentation-Based Context

In this section, segmentation-based methods are presented. Segmentation-based methods have been categorized into training-based and training-free approach. Each category is then subdivided into word image segmentation and text line segmentation context.

2.1. Methods Based on Training

2.1.1. Word Image Segmentation Context

In the work of Rodríguez-Serrano and Perronnin (2009), the method is based on a Semi-Continuous - Hidden Markov Model (SC-HMM) coupled with a Gaussian Mixture Model (GMM). SC-HMM is able to learn from a small set of samples. A segmentation algorithm extracts sub-images that potentially represent words, employing state-of-the-art techniques based on projection profiles and clustering of gap distances. Then, a simple classifier using holistic features (per column, pixel count, Local Gradient Histogram (LGH)) is employed for performing a first rejection pass. The non-rejected word images are normalized with respect to slant, skew and text height, using standard techniques. Then, for each normalized

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 308 of 404

word image, LGH features are computed by moving a window from left to right over the image and feature vectors are extracted at each position to build the feature vector sequence. Finally, using SC-HMM with GMM, a score is assigned to each feature vector sequence which is used to attribute the similarity with the query using a threshold. An overview of the methodology is shown in Figure 1.

The same framework was used in Rodríguez-Serrano et al. (2010) modified so that writers adaptation is achieved. For this purpose, a statistical adaptation technique was applied to change some of GMMs parameters at each document. Furthermore, SC-HMM was used in Rodríguez-Serrano and Perronnin (2012) to enrich features extraction, since in a left-to-right HMM, the states are ordered and the weights of the SC-HMMs can be viewed as a sequence of vectors. The distance between these vectors is computed using the Dynamic Time Warping (DTW) wherein Bhattacharyya measure is being used as local similarity. The use of SC-HMM in an unsupervised context was presented in Rodriguez-Serrano and Perronnin (2012) where character examples of existing fonts were used to create the training set.

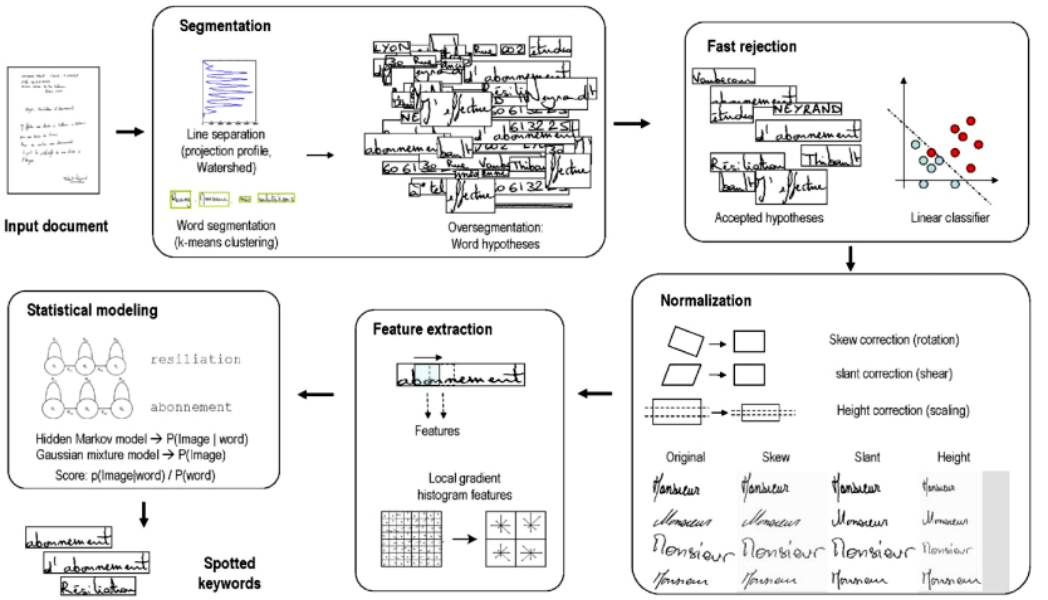


Figure 1. Overview of the word-spotting system presented in Rodríguez-Serrano and Perronnin (2009).

In the work in Almazán et al. (2012a), a preprocessing stage is initially applied using margins removal and anisotropic Gaussian filtering. Then, binarization and word segmentation are applied. In the core methodology, they use a hashing strategy based on Loci features to prune word images and limit the candidate locations. A discriminative model is then applied to those locations. The discriminative learning relies upon a Support Vector Machine (SVM) which sets the weights on the appearance features Histogram of Oriented Gradients (HOG) to compute the final similarity.

Almazan et al. (2013) created a method that is both QbE and QbS and addresses

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 309 of 404

multi-writer WS. The QbE pipeline is based on Fisher Vectors (FV) computed over Scale-Invariant Feature Transform SIFT descriptors. Then the FV are used to feed an SVM to get the attribute scores. They report that any encoding method and classification algorithm that transforms the input image into attribute scores could be used to replace them, but they chose SVMs and FVs for simplicity and effectiveness.

The study in Fernández et al. (2013) uses a previous method (Fernández et al., 2011) which is extended in a way that the syntactic context in the document is used to infer context. To achieve this, Markov Logic Networks (MLN) are used that are trained with specific rules. The MLN can be considered as a collection of first-order rules to each of which it is assigned a real number, the weight. Each rule represents the rule in the domain, while the weights indicate the strength of the rule.

The work in Aldavert et al. (2015) uses the Bag-of-Visual-Words (BoVW) model to WS. The authors divide the procedure of creating BoVW in four basic steps: sampling, description, encoding and pooling. In particular, they sampled densely the word images using a fixed step at different scales. The description is derived from the HOG descriptor. To encode the descriptors and create the codebook, the Locality-constrained Linear Coding (LLC) algorithm was used. Finally, at pooling step, the Spatial Pyramid Matching (SPM) technique applied so that spatial information could be used.

In the work at Sharma et al. (2015), they made experiments with Convolutional Neural Networks (CNN). Only the classification layers were retrained to address the problem. The CNN extracted a 4096-d feature for each word image which was achieved after discarding the last fully connected layer and considering the activation of the units in the second fully connected layer. For matching, standard Lp norms have been used.

2.1.2. Textline Segmentation Context

The system presented in Keaton et al. (1997) is composed of several modules. The “focus-of attention” module concerns a cross-correlation testing between the query image and the document for finding the candidate locations. The “preprocessing” module that consists of the estimation of word image zones at the upper, middle and lower level, filtering of stray marks and skeletonization. The “feature extraction” module that concerns profile encoding (20 Discrete Cosine Transform (DCT) coefficients from the profile extracted at each of the three zones and cavity encoding which takes into account 2D spatial arrangement, as well as the descender, and ascender information leading to a graph. Both encoding features are combined to a new graph that contains the type, size and relative location of each feature, which is considered as keyword signature graph. Finally, the keyword signature matching is addressed by two distinct comparisons. First, a comparison is employed between the profile encoding DCT coefficients wherein the resulting comparison is incorporated into the graph as an additional feature. In the sequel, the keyword signature graphs are compared with probabilistic graph matching based on Bayesian evidential reasoning.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 310 of 404

2.2. Training - Free Methods

2.2.1. Word Image Segmentation Context

In the paper Manmatha et al. (1996) the term "word spotting" was introduced for handwritten documents as analogous to "word spotting" in speech processing. It was applied on scanned gray level document images. The steps of the algorithm are gaussian filtering, subsampling to reduce the image by half, binarization by thresholding and segmentation into words. Then follows pruning taking into account (i) the aspect ratio and (ii) the size using predefined thresholds. Finally, at matching stage, two different matching algorithms were used based on the standard and the affine-corrected (SLH algorithm) Euclidean distance, respectively.

In the sequel, Rath and Manmatha (2003b), was motivated by Kolcz et al. (2000) that have used Dynamic Time Warping for matching in combination with a text line segmentation method. In this approach, the textline segmentation has been replaced by a word image segmentation. In the work of Rath and Manmatha (2003a), the feature set used for the experiments was extended to 11 distinct features (4 projection profiles, 2 word profiles, background ink transitions, gray level variance and 3 Gaussian smoothing variation features) which were used as single or combined ones for matching with DTW. Nevertheless, the more recent work in Rath and Manmatha (2007) suggested only three distinct features, namely, projection profiles, word profiles and background ink transitions which were optimally matched when using DTW.

A study based on contours of words was presented in Adamek et al. (2007). They start with local binarization of the image. To achieve smoothness at word outlines, they preprocess the image applying morphological filtering. After binarization, a word may split to more than one connected component. To estimate the exact position of the word in the word image a process based on horizontal and vertical projection histogram and a fixed thresholding is applied. Then, after applying a series of heuristics rules, the connected components of the word image are linked together to create a single component from which the contour is extracted. They used Multiscale Convexity Concavity (MCC) representation for the contour. MCC calculates the convexity and concavity along contour at different scales to create 2D matrix were rows correspond to scale level and columns to convexity or concavity. To measure the matching between the contours the DTW algorithm were used. The distance matrix of DTW is constructed by storing the distance between a pair of contour points corresponding to the row and column of the MCC representation. The final dissimilarity between contours is the normalized optimal path of the matrix. An alternative method that was tested is the MCC-DCT where the 1D DCT is applied at MCC representation matrix and the coefficients of DCT are combined to the final dissimilarity.

The work in Bhardwaj et al. (2008) presented an algorithm based on moment functions. In the initial stage, they used horizontal and vertical profiles to segment the document into lines and words, respectively. High order (up to 7) moment functions were used to extract features and indexing each word image. They used cosine similarity metric for matching and relevance feedback to improve the results.

A shape descriptor, the Compact Shape Portrayal Descriptor (CSPD) was presented in Zagoris et al. (2011) which requires only 123 bits per word image. CSPD is based on five (5) distinct features: (i) width height ratio, and the DCT coefficients of (ii) vertical projection,

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 311 of 404

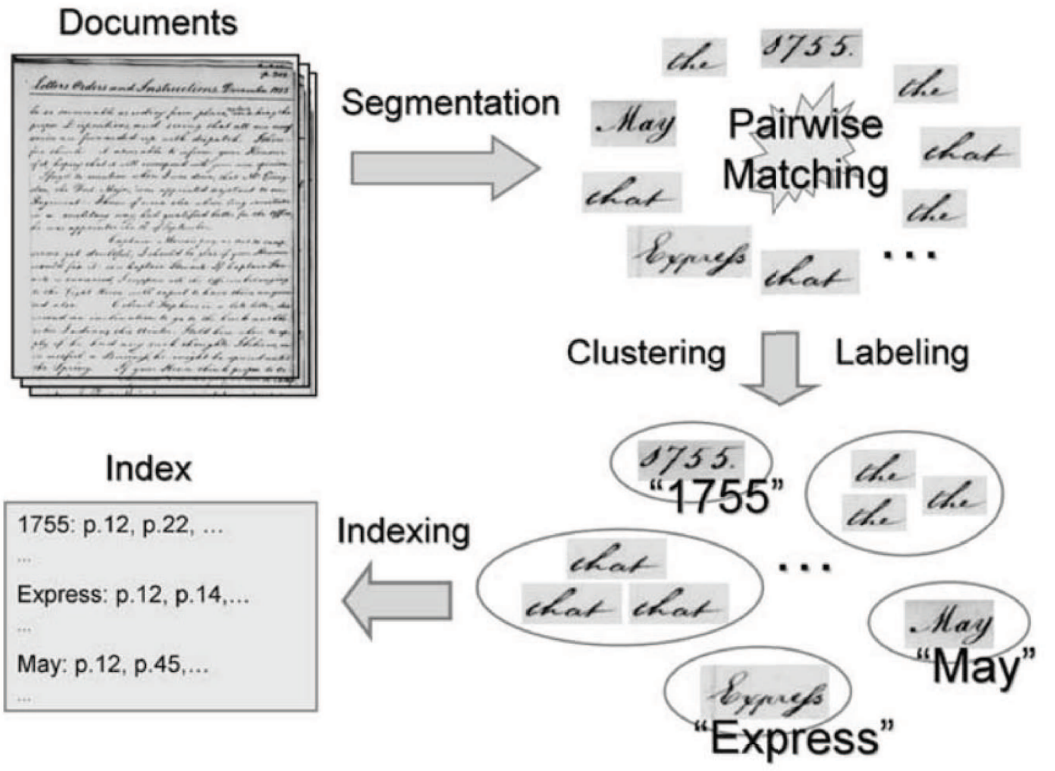


Figure 2. An illustration of the word spotting pipeline presented in (Rath and Manmatha, 2007).

(iii) horizontal projection, (iv) top shape projection and (v) bottom shape projection. For each feature, Gustafson-Kessel fuzzy algorithm is applied for quantization to reduce the stored size of each descriptor. To refine the results, relevance feedback is applied. The user selects the best result and a training set is created. This training set is used to train an SVM to correct the results. The similarity measure that was used is a modified (weighted) Minkowski L1.

The study appeared in Can and Duygulu (2011) was motivated from a shape representation method. They used already segmented word images provided with the datasets. The proposed methodology comprises 3 main steps. First, binarization is applied by using a threshold which is computed as the mean intensity value of the gray-scale image. Next, a contour extraction is used for each connected component in the binarized word image. Finally, a sequence of lines is created which is used as the descriptor for matching. The matching score is computed by first finding the distances between the line descriptors and then, summing all distances over the complete word image.

The basic premise in the work of Fornés et al. (2011) is that each word image is treated as a shape which can be represented using two models, namely the Blurred Shape Model (BSM) and the Deformable BSM. First, at preprocessing stage, segmented text lines are normalized by applying skew angle removal and slant correction. In the case of BSM,

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 312 of 404

the descriptor represents a probability distribution of the object shape. In the case of Deformable BSM, every image is represented with two output descriptors, a vector which contains the BSM value of each focus (equidistantly distributed points) and the position of each focus. The proposed matching technique lies upon the movement of focuses so that its own BSM is maximized. It is shown that using Euclidean distance in both BSM and deformable BSM methods outperforms the use of DTW.

Fernández et al. (2011) used Loci features (Glucksman, 1969) along eight directions. They are computed on the skeleton of each word image which is achieved after a document image binarization and word image segmentation step. The similarity is computed using the Euclidean and Cosine distance.

The study in Diesendruck et al. (2012b) and Diesendruck et al. (2012a) was focusing on building a search system for 1930-40 US Census data. The process starts with binarization, morphological thinning and Hough transform to locate table lines since the documents are in table format. Thus, the segmentation is based on table lines. Since each cell contains one word, the method is word-based segmented. Then, a signature vector was composed the first 10 coefficients of cosine transform of upper, lower and transition profiles. Since the data set is quite large and the response time should be reasonable, hierarchical agglomerative clustering with complete linkage is used to cluster the signature vectors.

The problem of sequential KWS was addressed in Fernández-Mota et al. (2014). In sequential KWS the ordered sequence of indices is taken into account for finding similar instances of words in a book. They experimented with descriptors that relate to a single writer scenario (BSM, HOG, nrHOG) as well as descriptors that relate to a multiple writers scenario (attribute-based approach).

The work in Howe (2013) is based on a part-structured modeling which aims to minimize a deformation energy required to fit the query to the word image instance. The process is initiated with a binarization. Then, skeletonization is applied to produce connected components of a single-pixel width. The endpoints and junctions of the skeleton are used to build a tree. An energy minimization of a function that comprises a deformation energy and an observation energy term is finally addressed.

The method in Kovalchuk et al. (2014) is the winner of H-KWS 2014 competition. First, they binarize the image by global thresholding and connected components are computed. Then, pruning of connected components is followed based on heuristics that rely upon properties of connected components. Using a regular grid of fixed size, they compute HOG and LBP descriptors which result in a 250D vector. A max-pooling process is then applied to the descriptor. The matching is made with L_2 distance.

In Wang et al. (2014) the authors initiate the process by applying the preprocessing step presented in Wang et al. (2013). They use a graph representation model which is based on the skeleton of each word image. In this graph, vertices are the structural interest points and the strokes connecting them are the edges. The value of vertices corresponds to the Shape Context descriptor while the value of edges corresponds to the length of the stroke. The computation of similarity between two-word images concerns the similarity of graphs for each connected component existing in the query and the test word image which is used to guide the DTW computation.

The work in Zagoris et al. (2014) is based on spatial information from word images. First, gradient vectors are calculated. Because of the sensitivity of gradient to noise, an Otsu

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 313 of 404

like threshold is applied to gradient vectors. At the remaining points, gradient orientation is calculated. Next, a linear quantization of gradients orientation to a desirable amount of levels follows. The quantization step also controls the amount of the final local points. After quantization, the corner points are characterized as initial keypoints (kP). The final points are the dominant kPs according to Shannon entropy at an area were the kP is the center. After the final points have been selected each area around kPs is divided into 9 subareas. For each of these 9 areas using a voting system based on the weighted distance of each point to kP, a 3-bin histogram is created. The combination of these 9 histograms to a 27-bin histogram results in the descriptor, the Document-Specific Local Features (DSLF). At matching stage first, a normalization is applied for each word image. Then instead of a brute force search, a Local Proximity Nearest Neighbour (LPNN) search is used by taking into account the mean distance of each pixel from the mean center. Finally, Euclidean distance is applied between the kPs and the results are presented in an ascending order. In Zagoris et al. (2015), an extension of this work is presented using Relevance Feedback strategies (CombSum, CombMin, Probabilistic model). It is reported that the optimal results are achieved from CombMin model.

The goal in Papandreou et al. (2016) is to study the zoning features. Binarization and deslanting are first applied to the query and candidate word images as pre-processing steps. The zoning features are extracted after cutting the query word in vertical zones based on its length and pixel distribution along the horizontal axis and adjusting these boundaries with the corresponding zones in the candidate word image using DTW. In the sequel, the word images are normalized and their features, which are based on pixel density, are extracted. Finally, the final distance is the product of Euclidean distance of the two-word images with the distance provided from DTW.

In Mondal et al. (2016) the author introduces a new matching technique the Flexible Sequence Matching (FSM) algorithm for KWS task. At the preprocessing stage, the document images are first binarized by an adaptive technique (Gatos et al., 2006), after binarization along with border removal is applied to obtain proper text boundaries (Stamatopoulos et al., 2010) and then a segmentation stage follows that partitions the documents in lines or pieces of lines, up to words or parts of words, depending on the experiment to be conducted. At features, extraction stage grayscale and binary image are used to extract two types of features, namely column-based features and Slit style HOG (SSHOG) (Terasawa and Tanaka, 2009). Eight Column-based features are extracted from the binary image. Finally, at matching stage, FSM is applied. FSM is similar in spirit to DTW but it is less sensitive to local variations in the spelling of words and to local degradations effects within the word image.

In a recent work Retsinas et al. (2016), three variances of the Projections of Oriented Gradient (POG) descriptor (Retsinas et al., 2015) studied in the framework of the KWS problem. The first variant, the global POG (gPOG) is slightly different from POG, for which the main differences are: (i) it keeps a different number of coefficients from DCT and (ii) has 6 projections. The second variant k-segmented POG (lPOG), first segments the word image to k overlapping images and then calculates the POG descriptor to each of them. The third variant, fusion POG (fPOG), as the name implies is a fusion of gPOG and lPOG descriptors. Finally, the Euclidean distance is used to attain the matching score. It should be noted that at the preprocessing step, binarization, skew correction and height normalization were applied.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 314 of 404

2.2.2. Textline Segmentation Context

In the paper Kolcz et al. (2000), the approach is motivated by the success of dynamic-programming-based techniques for KWS in speech applications even when very limited keyword models are present. It relies upon a line segmentation method to achieve distinct text lines for each document image. The ink-density histogram and its Fast Fourier Transform (FFT) spectrum are used to determine the text lines as well as the skew of the page. For each text line, they extract the upper and the lower profile as well as transitional Features (number of transactions between background and body in each column of pixels). Then they used DTW to address matching in the KWS pipeline. They used heuristics to reduce the computational time which was similar to the ones used by the work in (Manmatha et al., 1996).

Terasawa and Tanaka (2009) deals with language independent KWS method. They have chosen a line-oriented approach because (i) word segmentation is impossible in some languages and (ii) it can retrieve a hyphenated word that spans two lines. For each text line image, a narrow sliding window is used for feature extraction. A variation of HOG features is used, namely, the SSHOG features. Compared to the original HOG, the SSHOG uses a narrow window and the computed gradient is signed. For feature matching, DTW is used.

In Wang et al. (2013) the authors use a coarse-to-fine strategy. They first remove the noise with a smoothing filter and they apply text line segmentation based on Hough transform. At coarse step, they apply a sliding window at the size of the query word. They extract 4 textural features, namely, projection profile, upper and lower border and orientation distribution of skeleton pixels. Then they use DTW for the first three features and Chi-square metric for orientation distribution. With an empirical threshold, they chose the most similar to the query regions, the regions of interest. The fine step is applied to the regions of interest. In the fine step, morphological and topological properties are used. Morphological properties calculated using the Shape Context descriptor on selected interest points (branch points, starting/ending points and high-curved points). Topological properties are obtained from a skeletonized representation. The information of the properties is the input to a weighted distance function for which Linear Discriminant Analysis is used to automatically get the optimal weights.

3. Segmentation-Free Context

In this section, segmentation-free methods are presented. The methods of this section are divided into training-based and training-free.

3.1. Methods Based on Training

The approach in Choisy (2007) is made to deal with KWS of isolated words on mail envelopes. It is character segmentation-free which relies on the dynamic creation of global word models. This is achieved with the use of Non-Symmetric Half Plane - HMM (NSHP-HMM) (Saon and Belaïd, 1997) which is a model hybrid of an HMM and a Markov Field (MRF). Before applying NSHP-HMM, two preprocessing steps are applied (i) global slant correction and (ii) a non-linear normalization that centers and normalizes the lower case

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 315 of 404

zone of the writing. The NSHP-HMM is trained at the word level.

In the work of Rusinol et al. (2011), they used a BoVW model where each patch is normalized by applying the term frequency-inverse document frequency (tf-idf) model. BoVW words model powered by SIFT descriptors which were further augmented by a word segmentation. Then, Latent Semantic Indexing (LSI) (Deerwester et al., 1990) is used with BoVW to retrieve relevant patches even if they do not contain the same exact features than the query sub-image. Then Singular Value Decomposition (SVD) is applied to reduce the descriptor dimension and to obtain a transformed space where patches having similar topics but with different descriptors will lie close. At the retrieval stage, cosine similarity was used. In Rusinol et al. (2015), they have enhanced their preliminary version by including an indexation scheme aimed to scale the proposed method to handle large datasets. The SVD step replaced by Product Quantization (PQ) for this purpose. Also, a multi-length patch representation is also introduced, which increases the retrieval performance by taking into account the different possible lengths of the query words.

The study presented in Rusinol and Lladós (2012) comprises both fusion and relevance feedback mechanisms. At fusion stage, three different fusion methods were tested, namely, early fusion, combMAX and Borda count. For relevance feedback, three different methods were tested, the Aocchio's algorithm (Cao et al., 2011), the Ide dec-hi method (Ide, 1971) and the relevance score algorithm (Giacinto and Roli, 2004).

The work in Almazán et al. (2012b) is based on exemplar SVM for better representation of features. Documents are represented by a grid of HOG descriptors, and a sliding window approach is used to locate the document regions that are most similar to the query. They use the Exemplar SVM framework to produce a better representation of the query in an unsupervised way. Finally, the document descriptors are compressed with Product Quantization (PQ) which has also the benefit of calculating the distance between the query and the quantized document with the use of a look-up table.

The method presented in Dovgalecs et al. (2013) may operate with words or graphics and is situated in the BoVW framework. In the offline stage, the BoVW is created from densely sampled SIFT features. In the online stage, candidate zones detection works by comparing query features with BoVW using chi-square distance. Then, the Longest Weighted Profile (LWP) algorithm which enforces spatial ordering information characteristics of words and graphical patterns alike, is used to compute the similarity score between the query image and the candidate zones.

The study in Rothacker et al. (2013) is based on the use of Bag-of-Features with HMMs. The method is divided into three parts. First, densely SIFT features are extracted in the whole document image and 5% of those are used to create a codebook size of 4096 for Bag-of-Features. Then, the Bag-of-Features representation feeds an HMM which encodes the sequential visual appearance of features that are located in the query bounding box. Finally, the document collection can be queried in a patch-based fashion where the output is a map of probabilistic scores from which the query results can be retrieved. As shown in the upper part of Figure 3 the document image representation is visualized. In the lower part, the estimation of a query model and the patch-based decoding with respect to that model are shown. Patch-based representations are evaluated at each grid point. The scores obtained are visualized by interpolating them over the document image indicating low to high responses with blue to red colors.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 316 of 404

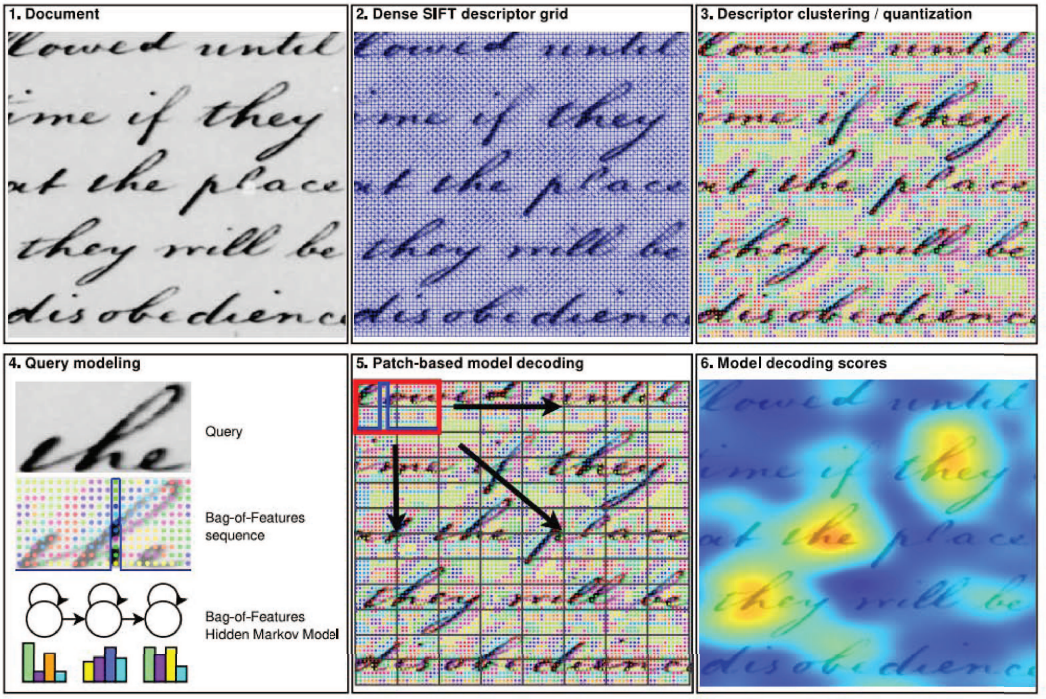


Figure 3. Overview of the segmentation-free word spotting method presented in (Rothacker et al., 2013).

3.2. Training - Free Methods

The work in Leydier et al. (2009) focuses on medieval Latin manuscripts and is based on the observation that medieval Latin manuscripts have letters mainly composed of large vertical strokes. The algorithm has two main steps. In the first step, guides, gradients and Zones Of Interest (ZOI)s are extracted from the document image and the query word image. In the second step, cohesive matching is applied between the guides of the query image and those of the document image. For each match of guides, a check if ZOIs are matching is applied, too. This work was enriched with a model that is the combination of an alphabet, a glyph book and a grammar as presented in Leydier et al. (2009). The model is used to create a character tree. The extra information from the character tree was used among the query word image and the document image for better extraction of ZOIs, guides and gradients. They also automated thresholds that were needed at ZOIs, guides and gradients extraction, but also at cohesive matching.

Zhang and Tan (2013) is motivated by the Heat Kernel Signature (HKS) which has been used for shape recognition. Actually, the Deformation and Light Invariant (DaLI) descriptor was used which is applied by convolving Scale Invariant HKS (SI-HKS) with Gaussian kernels. To compute the similarity, a Delaunay Triangulation algorithm was applied to create a Triangular Mesh Structure (TMS) of keypoints detected in the word image and the document image, respectively. Finally, the similarity score is computed by building a score matrix which contains the optimal matching score and the optimal matching history.

The approach in Hast and Fornés (2016) is based on Putative Match Analysis (PUMA)

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 317 of 404

(Hast and Marchetti, 2012), a technique that first introduced for matching aerial images. First, the input images are binarized using Otsu and then smoothed with a Gaussian in order to find more key points. Then, four different kinds of key points are detected in the word images, which basically detect lines, corners and blobs. Taking into account the detected keypoints, Fourier-based feature descriptors are computed. At the end, the matching is performed by an improved version of Random Sample Consensus (RANSAC), called as PUMA, which is able to allow a more relaxed matching among the word images.

Rabaev et al. (2016) focus on KWS by locating the query word in a document recursively in a scale-space pyramid. The proposed scheme does not depend on a specific choice of features. They experimented with the HOG descriptors, which have been shown to provide good results. Chi-square distance is applied to compare HOG descriptors at all levels of the pyramid.

4. Experimental Datasets and Evaluation Metrics

4.1. Evaluation Metrics

The evaluation metrics that have been used for performance evaluation between different word spotting algorithms are inspired from information retrieval. Therefore, each retrieved item (word) is defined as relevant to the original query (word - query) or not. Early reports on the KWS performance evaluation were simply taking the first n words and calculate the most basic retrieval metrics, the **Precision** and **Recall**.

$$Precision = \frac{\{relevant\ words\} \cap \{retrieved\ words\}}{\{retrieved\ words\}} \tag{1}$$

$$Recall = \frac{\{relevant\ words\} \cap \{retrieved\ words\}}{\{relevant\ words\}} \tag{2}$$

Precision is the fraction of retrieved words that are relevant to the search, while Recall is the fraction of the words that are relevant to the query. It is apparent that the above metrics are inversely related. To achieve a combined evaluation, the precision-recall curve is computed.

The **Precision - Recall Curve** is computed by the traditional 11-point interpolated average precision approach (Manning et al., 2008), (Van Rijsbergen, 1979). For each query, the interpolated precision is measured at the 11 recall levels of 0.0, 0.1, 0.2, ..., 1.0.

Sometimes, the differences between the evaluation algorithms are very hard to observe especially, between very small performance results. Moreover, these graphs may not contain all the desired information (Salton, 1992). Therefore, the need to evaluate the retrieval results with a single value is needed. The most common evaluation metric that can meet this requirement is the **Mean Average Precision (MAP)** (NIST, 2013; Chatzichristofis et al., 2011) which is defined as the average of the precision value obtained after each relevant word is retrieved:

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 318 of 404

Table 1. Descriptors, learning methods and similarity measures used from each method

	Method	Descriptors	Learning	Similarity
Training based	(Rodríguez-Serrano and Perronnin, 2009)	LGH	SC-HMM	Euclidean
	(Rodríguez-Serrano and Perronnin, 2012)	LGH	SC-HMM	DTW
	(Almazán et al., 2012a)	Loci, HOG	SVM	Dot product
	(Fernández et al., 2013)	Loci	MLN	Euclidean
	(Aldavert et al., 2015)	HOG	BoVW	Histogram Matching
	(Sharma et al., 2015)	Deep features	CNN	Lp-norm
	(Keaton et al., 1997)	DCT on profiles	Bayesian network	Graph matching
	(Choisy, 2007)	Column-wise binary patterns	NSHP-HMM	Posteriori Probability
	(Rusinol et al., 2011)	SIFT	BoVW	Histogram Matching
	(Almazán et al., 2012b)	HOG	Exemplar SVM	Euclidean
	(Dovgalecs et al., 2013)	SIFT	BoVW	Chi-square
	(Rothacker et al., 2013)	SIFT	BoVW, HMM	Histogram Matching
Training free	(Manmatha et al., 1996)	Profiles		Euclidean, DTW
	(Adamek et al., 2007)	MCC, DCT		DTW
	(Bhardwaj et al., 2008)	Moments		Cosine
	(Zagoris et al., 2011)	CSPD		Minkowski L1
	(Can and Duygulu, 2011)	Sequence of lines		Line matching
	(Fornés et al., 2011)	Deformable BSM		Euclidean, DTW
	(Fernández et al., 2011)	Loci		Euclidean, Cosine
	(Diesendruck et al., 2012b,a)	DCT on profiles		Euclidean
	(Howe, 2013)	Endpoints and junctions of skeleton		Energy minimization
	(Kovalchuk et al., 2014)	HOG, LBP		Euclidean
	(Wang et al., 2014)	SC		DTW
	(Zagoris et al., 2014)	DSLF		Euclidean
	(Papandreou et al., 2016)	Zoning features		Euclidean and DTW
	(Mondal et al., 2016)	Column-based, SSHOG		FSM
	(Retsinas et al., 2016)	POG, gPOG, lPOG, fPOG		Euclidean
	(Kolcz et al., 2000)	Profiles		DTW
	(Terasawa and Tanaka, 2009)	SSHOG		DTW
	(Wang et al., 2013)	Profiles, SC		DTW
	(Leydier et al., 2009)	ZOI		Cohesive matching
	(Zhang and Tan, 2013)	DaLI		Minimum cost path between connected keypoints in a mesh grid
	(Hast and Fornés, 2016)	Corners, blobs		PUMA
	(Rabaev et al., 2016)	HOG		Chi-square

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 319 of 404

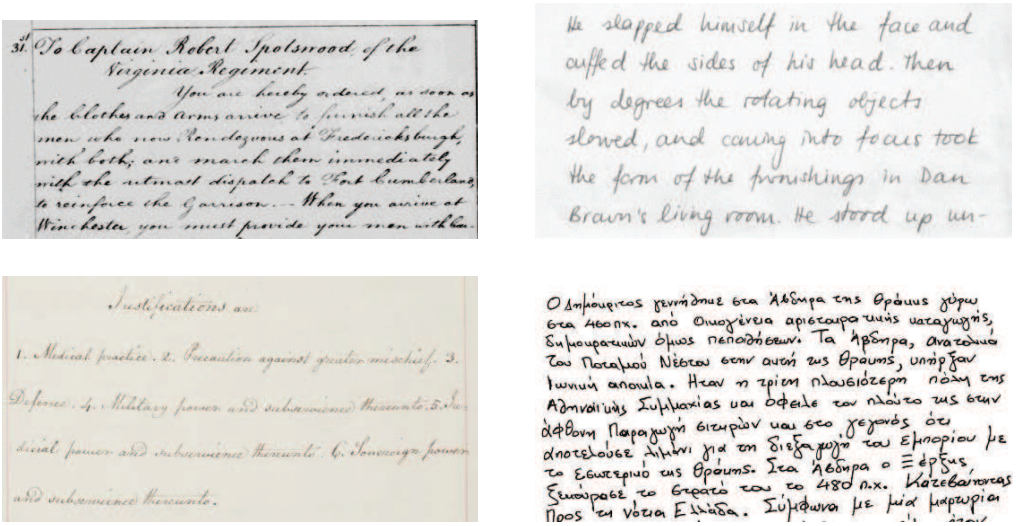


Figure 4. Samples from the most used datasets (a) GW (b) IAM (c) Bentham (d) Modern.

$$AP = \frac{\sum_{k=1}^{n} (P@K \times rel(k))}{\{relevant\ words\}} \tag{3}$$

where

Precision at k items (P@k) denoted as

$$P@k = \frac{\{relevant\ words\} \cap \{k\ retrieved\ words\}}{\{k\ retrieved\ words\}} \tag{4}$$

with the function of relevance denoted as follows:

$$rel(k) = \begin{cases} 1 & \text{if word at rank } k \text{ is relevant} \\ 0 & \text{if word at rank } k \text{ is not relevant} \end{cases}$$

Finally, the **Mean Average Precision (MAP)** is calculated by averaging the AP for all the queries, denoted as:

$$MAP = \sum_{q=1}^{Q} AP_q \tag{5}$$

where Q is the total number of queries.

It is worth to note that in the experiments for the segmentation-free case, a resulting bounding box may not match exactly with the word bounding box from ground-truth corpora. Thus, a correct match is registered when the relative overlapping area is over a certain threshold. For the sake of consistency, in every segmentation-free experiment the overlapping area is defined as:

$$OA = \frac{A \cap B}{A \cup B} \tag{6}$$

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 320 of 404

where *OA* is the overlapping area, *A* the resulting bounding box and *B* is the ground-truth.

The challenging nature of KWS in handwritten documents has motivated the organization of three dedicated international competitions in conjunction with the International Conference on Frontiers of Handwriting Recognition (ICFHR) and the International Conference on Document Analysis and Recognition (ICDAR). In particular, the ICFHR 2014 Handwritten Keyword Spotting Competition (ICFHR-2014) (Pratikakis et al., 2014), the ICDAR 2015 Competition on Keyword Spotting for Handwritten Documents (ICDAR-2015) (Puigcerver et al., 2015) and the ICFHR 2016 Handwritten Keyword Spotting Competition (ICFHR-2016) (Pratikakis et al., 2016) have been the venues where research groups have competed in two different KWS scenarios, namely, segmentation-free and segmentation-based. Table 2, 3 and 4 shows the results for the ICFHR-2014, ICDAR-2015 and ICFHR-2016 competitions, respectively.

Table 2. Experimental results for the ICFHR-2014 Competition

	Segmentation-based				Segmentation-free			
	BENTHAM		MODERN		BENTHAM		MODERN	
Method	mAP	P@5	mAP	P@5	mAP	P@5	mAP	P@5
(Kovalchuk et al., 2014)	0.524	0.738	0.338	0.588	0.416	0.609	0.263	0.539
(Almazan et al., 2013)	0.513	0.724	0.523	0.706	-	-	-	-
(Howe, 2013)	0.462	0.718	0.278	0.569	0.363	.556	0.163	0.417
(Leydier et al., 2009)	-	-	-	-	0.205	0.335	0.087	0.234
(Pantke et al., 2013)	-	-	-	-	0.337	0.543	0.091	0.245

Table 3. Experimental results for the ICDAR-2015 Competition

	Segmentation-based		Segmentation-free	
Method	mAP	P@5	mAP	P@5
Pattern Recognition Group, TU Dortmund University	0.424	0.406	0.276	0.343
(Aldavert et al., 2013)	0.300	0.342	0.082	0.109

Table 4. Experimental results for the ICFHR-2016 Competition

	Segmentation-based		Segmentation-free	
	Botany	Konzil.	Botany	Konzil.
Method	mAP	mAP	mAP	mAP
Computer Vision Center (CVCDAG) Universitat Autoònoma de Barcelona, Spain	75.77	77.91	0	0
Pattern Recognition (PRG) TU Dortmund University, Germany	89.69	96.05	15.89	52.20
(Kovalchuk et al., 2014)	50.64	71.11	37.48	61.78
Visual Information and Interaction (QTOB) Uppsala University, Sweeden	54.95	82.15	-	-

"CB" stands for a collection of 50 pages from handwritten marriage licenses from the Barcelona Cathedral written in 1617. URL: http://dag.cvc.uab.es/the-esposalles-database/

The "Bentham" dataset is part of the H-KWS 2014 contest's dataset. It consists of high

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 321 of 404

Table 5. Segmentation type and Datasets used from each method.

	Method	Segmentation type		Dataset
		Line	Word	
Training based	(Rodríguez-Serrano and Perronnin, 2009), (Rodriguez-Serrano and Perronnin, 2012), (Rodríguez-Serrano et al., 2010)		X	French
	(Rodríguez-Serrano and Perronnin, 2012)		X	GW, French, IFN/ENIT
	(Almazán et al., 2012a), (Almazan et al., 2013)		X	CB
	(Fernández et al., 2013)		X	CB
	(Aldavert et al., 2015)		X	GW, Bentham, Modern
	(Sharma et al., 2015)		X	IAM
	(Keaton et al., 1997)	X		AIS
	(Choisy, 2007)	-	-	HMS
	(Rusinol et al., 2011, 2015)	-	-	GW
	(Almazán et al., 2012b)	-	-	GW
	(Dovgalecs et al., 2013)	-	-	GW
	(Rothacker et al., 2013)	-	-	GW
Training free	(Manmatha et al., 1996)		X	DIMUND, Hudson
	(Rath and Manmatha, 2003a,b, 2007)		X	GW
	(Adamek et al., 2007)		X	GW
	(Bhardwaj et al., 2008)		X	GW, IAM
	(Zagoris et al., 2011)		X	GW, Greek
	(Can and Duygulu, 2011)		X	GW, OTM
	(Fornés et al., 2011)		X	GW
	(Fernández et al., 2011), (Fernández-Mota et al., 2014)		X	CB
	(Diesendruck et al., 2012b,a)		X	USC
	(Howe, 2013)		X	GW
	(Kovalchuk et al., 2014)		X	GW
	(Wang et al., 2014)		X	GW, CB
	(Zagoris et al., 2014)		X	GW, Bentham, Modern
	(Papandreou et al., 2016)		X	GW, Bentham
	(Mondal et al., 2016)		X	GW, Japanese
	(Retsinas et al., 2016)		X	Bentham, Modern
	(Kolcz et al., 2000)	X		AIS
	(Terasawa and Tanaka, 2009)	X		GW, Japanese
	(Wang et al., 2013)	X		CITERE
	(Leydier et al., 2007, 2009)	-	-	GW
	(Zhang and Tan, 2013)	-	-	GW
	(Hast and Fornés, 2016)	-	-	CB
	(Rabaev et al., 2016)	-	-	GW, CG, Arabic

quality (approximately 3000 pixels width and 4000 pixels height) handwritten manuscripts. The documents are written by Jeremy Bentham (1748-1832) himself as well as by Bentham's secretarial staff over a period of sixty years.

The "Modern" dataset is also part of the H-KWS 2014 contest's dataset. It consists of modern handwritten documents from the ICDAR 2009 Handwritten Segmentation Contest. These documents originate from several writers that were asked to copy a given text. They

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 322 of 404

do not include any non-text elements (lines, drawings, etc.) and are written in four (4) languages (English, French, German and Greek).

A dataset that comprises 1539 pages of modern off-line handwritten English text written by 657 different writers is denoted as “IAM”. URL: http://www.fki.inf.unibe.ch/databases/iam-handwriting-database

The Archives of the Indies in Seville (AIS), is a repository that represents the official communication between the Spanish Crown and its New World colonies and spans approximately four centuries (i.e. 15th-19th).

At Choisy (2007) a dataset that consists of French handwritten mails collection was used for Handwritten Mail Sorting (HMS) task, wherein 1522 mail pages are manually labelled.

At Manmatha et al. (1996) as dataset two single pages were used. One obtained from the DIMUND document server, thus denoted as “DIMUND” and the other single page was taken from a collection in the library of the University of Massachusetts. This page is a letter written by James S. Gibbons to Erasmus Darwin Hudson and it is denoted as “Hudson”.

An Ottoman dataset denoted as “OTM” comprises documents written with a commonly encountered calligraphy style called Riqqa, which was used in official documents. Consists of 257 words in three pages of text. URL: http://courses.washington.edu/otap/

US Census forms from 1930 and 1940 comprise a dataset denoted as “USC”.

Scanned images of the Japanese manuscript “The diary of Matsumae Kageyu” by Akoku Raishiki comprise a dataset denoted as “Japanese”.

Letters written by different French philosophers constituting 4 collections are denoted as “CITERE”. There are 11 pages containing approximately 2000 words, where 51 words were used as queries. URL: http://citere.hypotheses.org/

The Cairo Geniza (CG) collection that consists of 12 document images dated to the 10th century. This collection exhibits smeared characters, bleed through, and stains. The page size is about 1650×2330 pixels, and the collection contains 1371 words of 921 different transcriptions. URL: http://www.genizah.org/

A collection of 10 pages of Islamic manuscripts from Harvard University denoted as “Arabic” consists of documents that are dated from 12th to 15th centuries. The ground truth for this collection is given in terms of word-parts. Since word-parts are relatively small, for the experiments, 5117 largest word-parts were chosen with 929 different transcriptions. The page size is 1600 pixels. URL: http://ocp.hul.harvard.edu/ihp/.

Conclusion

The major difference between segmentation-based and segmentation-free methods is the different search space (distinct word images versus the whole document image) where they operate. This is the basis of each disadvantage or advantage that each approach entails.

The main advantage of the segmentation-based methods is the retrieval speed. The ability to know the words boundaries inside the document provide profound advantages with respect to an efficient retrieval performance. Therefore, segmentation-based methods are mainly based on word segmentations as there is only one method i.e. Keaton et al. (1997) which alternatively uses text line segmentation.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 323 of 404

On the other hand, if the document is very noisy or very complex to apply a word segmentation method then a segmentation-free methodology is the most appropriate approach. Unfortunately, current segmentation-free KWS methods seem not to be appealing since the indexing storage size and the retrieval computation are very costly even when they are dealing with a medium size (100 pages) datasets. Moreover, the complicated issue of managing directly the whole document image is the main reason that few works deal with the segmentation-free approach.

Concerning features extraction, it is worth noting, that the majority of the works are relying on some form of gradients like SIFT, HOG, LGH, SSHOG, POG. Although the initial approaches were using profile features, recent works understood the spatial texture information is more robust than shape, especially for the handwritten documents. Lastly, some very recent works use features that are obtained using Deep Neural Networks which are called deep features.

For the training-based methods, the BoVW has been extensively used as a standalone learning component or combined with other models like the HMMs. The connection with the HMMs was motivated by the use of HMMs for handwritten transcription using a modeling inherent to the way a human makes a transcription. The recent advent of CNNs has started to appear in the KWS context (Sharma et al., 2015).

Until recently, the most common used dataset was the George Washington dataset for which, there was not a common evaluation protocol and each researcher employing its own subset and query set. Fortunately, the recent KWS competitions have set the ground for a concise performance evaluation framework.

Finally, in some works (Zagoris et al., 2011, 2014; Rusinol et al., 2011; Bhardwaj et al., 2008) the KWS pipeline is coupled with a relevance feedback mechanism which introduces the user in the retrieval loop, thus, improving the final retrieval performance.

References

Adamek, T., O'Connor, N. E., and Smeaton, A. F. (2007). Word matching using single closed contours for indexing handwritten historical documents. *International Journal of Document Analysis and Recognition (IJDAR)*, 9(2-4):153–165.

Aldavert, D., Rusinol, M., Toledo, R., and Lladós, J. (2013). Integrating visual and textual cues for query-by-string word spotting. In Document Analysis and Recognition (ICDAR), 2013 12th International Conference on, pages 511–515. IEEE.

Aldavert, D., Rusiñol, M., Toledo, R., and Lladós, J. (2015). A study of bag-of-visual-words representations for handwritten keyword spotting. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(3):223–234.

Almazán, J., Fernández, D., Fornés, A., Lladós, J., and Valveny, E. (2012a). A coarse-to-fine approach for handwritten word spotting in large scale historical documents collection. In Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on, pages 455–460. IEEE.

Almazán, J., Gordo, A., Fornés, A., and Valveny, E. (2012b). Efficient exemplar word spotting. In *BMVC*, volume 1, page 3.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 324 of 404

Almazan, J., Gordo, A., Fornés, A., and Valveny, E. (2013). Handwritten word spotting with corrected attributes. In Proceedings of the IEEE International Conference on Computer Vision, pages 1017–1024.

Bhardwaj, A., Jose, D., and Govindaraju, V. (2008). Script independent word spotting in multilingual documents. In *IJCNLP*, pages 48–54.

Can, E. F. and Duygulu, P. (2011). A line-based representation for matching words in historical manuscripts. *Pattern Recognition Letters*, 32(8):1126–1138.

Cao, H., Govindaraju, V., and Bhardwaj, A. (2011). Unconstrained handwritten document retrieval. *International Journal on Document Analysis and Recognition (IJDAR)*, 14(2):145–157.

Chatzichristofis, S. A., Zagoris, K., and Arampatzis, A. (2011). The TREC files: the (ground) truth is out there. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pages 1289–1290. ACM.

Choisy, C. (2007). Dynamic handwritten keyword spotting based on the NSHP-HMM. In Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on, volume 1, pages 242–246. IEEE.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Diesendruck, L., Marini, L., Kooper, R., Kejriwal, M., and McHenry, K. (2012a). Digitization and search: A non-traditional use of hpc. In E-Science (e-Science), 2012 IEEE 8th International Conference on, pages 1–6. IEEE.

Diesendruck, L., Marini, L., Kooper, R., Kejriwal, M., and McHenry, K. (2012b). A framework to access handwritten information within large digitized paper collections. In E-Science (e-Science), 2012 IEEE 8th International Conference on, pages 1–10. IEEE.

Dovgalecs, V., Burnett, A., Tranouez, P., Nicolas, S., and Heutte, L. (2013). Spot it! finding words and patterns in historical documents. In Document Analysis and Recognition (ICDAR), 2013 12th International Conference on, pages 1039–1043. IEEE.

Fernández, D., Lladós, J., and Fornés, A. (2011). Handwritten word spotting in old manuscript images using a pseudo-structural descriptor organized in a hash structure. In Iberian Conference on Pattern Recognition and Image Analysis, pages 628–635. Springer.

Fernández, D., Marinai, S., Lladós, J., and Fornés, A. (2013). Contextual word spotting in historical manuscripts using markov logic networks. In Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing, pages 36–43. ACM.

Fernández-Mota, D., Manmatha, R., Fornes, A., and Llados, J. (2014). Sequential word spotting in historical handwritten documents. In Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on, pages 101–105. IEEE.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 325 of 404

Fornés, A., Frinken, V., Fischer, A., Almazán, J., Jackson, G., and Bunke, H. (2011). A keyword spotting approach using blurred shape model-based descriptors. In Proceedings of the 2011 workshop on historical document imaging and processing, pages 83–90. ACM.

Gatos, B., Pratikakis, I., and Perantonis, S. J. (2006). Adaptive degraded document image binarization. *Pattern recognition*, 39(3):317–327.

Giacinto, G. and Roli, F. (2004). Instance-based relevance feedback for image retrieval. In *NIPS*, pages 489–496.

Glucksman, H. A. (1969). Classification of mixed-font alphabetics by characteristic loci. Technical report, DTIC Document.

Hast, A. and Fornés, A. (2016). A segmentation-free handwritten word spotting approach by relaxed feature matching. In Document Analysis Systems (DAS), 2016 12th IAPR Workshop on, pages 150–155. IEEE.

Hast, A. and Marchetti, A. (2012). Putative match analysis: a repeatable alternative to RANSAC for matching of aerial images. In International Conference on Computer Vision Theory and Applications, VISAPP2012, Rome, Italy, 24-26 February, 2012, pages 341–344. SciTePress.

Howe, N. R. (2013). Part-structured inkball models for one-shot handwritten word spotting. In Document Analysis and Recognition (ICDAR), 2013 12th International Conference on, pages 582–586. IEEE.

Ide, E. (1971). New experiments in relevance feedback. *The SMART retrieval system*, pages 337–354.

Keaton, P., Greenspan, H., and Goodman, R. (1997). Keyword spotting for cursive document retrieval. In Document Image Analysis, 1997.(DIA'97) Proceedings., Workshop on, pages 74–81. IEEE.

Kolcz, A., Alspector, J., Augusteijn, M., Carlson, R., and Popescu, G. V. (2000). A line-oriented approach to word spotting in handwritten documents. *Pattern Analysis & Applications*, 3(2):153–168.

Kovalchuk, A., Wolf, L., and Dershowitz, N. (2014). A simple and fast word spotting method. In Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on, pages 3–8. IEEE.

Leydier, Y., Lebourgeois, F., and Emptoz, H. (2007). Text search for medieval manuscript images. *Pattern Recognition*, 40(12):3552–3567.

Leydier, Y., Ouji, A., LeBourgeois, F., and Emptoz, H. (2009). Towards an omnilingual word retrieval system for ancient manuscripts. *Pattern Recognition*, 42(9):2089–2105.

Manmatha, R., Han, C., and Riseman, E. M. (1996). Word spotting: A new approach to indexing handwriting. In Computer Vision and Pattern Recognition, 1996. Proceedings CVPR'96, 1996 IEEE Computer Society Conference on, pages 631–637. IEEE.

Samsung Ex. 1011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 326 of 404

Manning, C. D., Raghavan, P., Schütze, H., et al. (2008). *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge.

Mondal, T., Ragot, N., Ramel, J.-Y., and Pal, U. (2016). Flexible sequence matching technique: An effective learning-free approach for word spotting. *Pattern Recognition*, 60:596–612.

NIST, T. (2013). TREC NIST. [Online]. Available: http://trec.nist.gov/pubs/trec16/appendices/measures.pdf.

Pantke, W., Märgner, V., Fecker, D., Fingscheidt, T., Asi, A., Biller, O., El-Sana, J., Saabni, R., and Yehia, M. (2013). Hadara–a software system for semi-automatic processing of historical handwritten arabic documents. In Archiving Conference, volume 2013, pages 161–166. Society for Imaging Science and Technology.

Papandreou, A., Gatos, B., and Zagoris, K. (2016). An adaptive zoning technique for word spotting using dynamic time warping. In Document Analysis Systems (DAS), 2016 12th IAPR Workshop on, pages 387–392. IEEE.

Pratikakis, I., Zagoris, K., Gatos, B., Louloudis, G., and Stamatopoulos, N. (2014). ICFHR 2014 competition on handwritten keyword spotting (H-KWS 2014). In Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on, pages 814–819. IEEE.

Pratikakis, I., Zagoris, K., Gatos, B., Puigcerver, J., Toselli, A. H., and Vidal, E. (2016). ICFHR 2016 handwritten keyword spotting competition (h-kws 2016). In Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on, pages 613–618. IEEE.

Puigcerver, J., Toselli, A. H., and Vidal, E. (2015). ICDAR 2015 competition on keyword spotting for handwritten documents. In Document Analysis and Recognition (ICDAR), 2015 13th International Conference on, pages 1176–1180. IEEE.

Rabaev, I., Kedem, K., and El-Sana, J. (2016). Keyword retrieval using scale-space pyramid. In Document Analysis Systems (DAS), 2016 12th IAPR Workshop on, pages 144–149. IEEE.

Rath, T. M. and Manmatha, R. (2003a). Features for word spotting in historical manuscripts. In Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on, pages 218–222. IEEE.

Rath, T. M. and Manmatha, R. (2003b). Word image matching using dynamic time warping. In Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, volume 2, pages II–II. IEEE.

Rath, T. M. and Manmatha, R. (2007). Word spotting for historical documents. *International Journal of Document Analysis and Recognition (IJDAR)*, 9(2-4):139–152.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 327 of 404

Retsinas, G., Gatos, B., Stamatopoulos, N., and Louloudis, G. (2015). Isolated character recognition using projections of oriented gradients. In Document Analysis and Recognition (ICDAR), 2015 13th International Conference on, pages 336–340. IEEE.

Retsinas, G., Louloudis, G., Stamatopoulos, N., and Gatos, B. (2016). Keyword spotting in handwritten documents using projections of oriented gradients. In Document Analysis Systems (DAS), 2016 12th IAPR Workshop on, pages 411–416. IEEE.

Rodríguez-Serrano, J. A. and Perronnin, F. (2009). Handwritten word-spotting using hidden markov models and universal vocabularies. *Pattern Recognition*, 42(9):2106–2116.

Rodríguez-Serrano, J. A. and Perronnin, F. (2012). A model-based sequence similarity with application to handwritten word spotting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2108–2120.

Rodriguez-Serrano, J. A. and Perronnin, F. (2012). Synthesizing queries for handwritten word image retrieval. *Pattern Recognition*, 45(9):3270–3276.

Rodríguez-Serrano, J. A., Perronnin, F., Sánchez, G., and Lladós, J. (2010). Unsupervised writer adaptation of whole-word HMMs with application to word-spotting. *Pattern Recognition Letters*, 31(8):742–749.

Rothacker, L., Rusinol, M., and Fink, G. A. (2013). Bag-of-features HMMs for segmentation-free word spotting in handwritten documents. In Document Analysis and Recognition (ICDAR), 2013 12th International Conference on, pages 1305–1309. IEEE.

Rusinol, M., Aldavert, D., Toledo, R., and Llados, J. (2011). Browsing heterogeneous document collections by a segmentation-free word spotting method. In Document Analysis and Recognition (ICDAR), 2011 International Conference on, pages 63–67. IEEE.

Rusinol, M., Aldavert, D., Toledo, R., and Lladós, J. (2015). Efficient segmentation-free keyword spotting in historical document collections. *Pattern Recognition*, 48(2):545–555.

Rusinol, M. and Lladós, J. (2012). The role of the users in handwritten word spotting applications: query fusion and relevance feedback. In Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on, pages 55–60. IEEE.

Salton, G. (1992). The state of retrieval system evaluation. *Information processing & management*, 28(4):441–449.

Saon, G. and Belaïd, A. (1997). High performance unconstrained word recognition system combining HMMs and markov random fields. *International Journal of Pattern Recognition and Artificial Intelligence*, 11(05):771–788.

Sharma, A. et al. (2015). Adapting off-the-shelf CNNs for word spotting & recognition. In Document Analysis and Recognition (ICDAR), 2015 13th International Conference on, pages 986–990. IEEE.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 328 of 404

Stamatopoulos, N., Gatos, B., and Georgiou, T. (2010). Page frame detection for double page document images. In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, pages 401–408. ACM.

Terasawa, K. and Tanaka, Y. (2009). Slit style hog feature for document image word spotting. In Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on, pages 116–120. IEEE.

Van Rijsbergen, C. (1979). Information retrieval, 2nd Edition, Butterworths.

Wang, P., Eglin, V., Garcia, C., Largeron, C., Lladós, J., and Fornés, A. (2014). A novel learning-free word spotting approach based on graph representation. In Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on, pages 207–211. IEEE.

Wang, P., Eglin, V., Garcia, C., Largeron, C., and McKenna, A. (2013). A comprehensive representation model for handwriting dedicated to word spotting. In Document Analysis and Recognition (ICDAR), 2013 12th International Conference on, pages 450–454. IEEE.

Zagoris, K., Ergina, K., and Papamarkos, N. (2011). Image retrieval systems based on compact shape descriptor and relevance feedback information. *Journal of Visual Communication and Image Representation*, 22(5):378–390.

Zagoris, K., Pratikakis, I., and Gatos, B. (2014). Segmentation-based historical handwritten word spotting using document-specific local features. In Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on, pages 9–14. IEEE.

Zagoris, K., Pratikakis, I., and Gatos, B. (2015). A framework for efficient transcription of historical documents using keyword spotting. In Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing, pages 9–14. ACM.

Zhang, X. and Tan, C. L. (2013). Segmentation-free keyword spotting for handwritten documents based on heat kernel signature. In Document Analysis and Recognition (ICDAR), 2013 12th International Conference on, pages 827–831. IEEE.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 329 of 404

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

In: Handwriting: Recognition, Development and Analysis ISBN: 978-1-53611-937-4
Editors: Byron L. D. Bezerra et al. © 2017 Nova Science Publishers, Inc.

Chapter 12

HANDWRITING-ENABLED E-PAPER BASED ON TWISTING-BALL DISPLAY

Yusuke Komazaki* *and Toru Torii*
Graduate School of Frontair Sciences,
The University of Tokyo, Chiba, Japan

1. Introduction

Recently, electronic paper (e-paper) has attracted increasing attention as a reflective and low-power display device and it has come into use in many fields such as e-readers, price tags, digital signage, light modulation and even fashions. Although electrophoretic display(Ota et al., 1973; Jacobson et al., 1998) is the most famous display known as e-paper, this is a generic term referring to thin and lightweight reflective displays and there are several types of displays classified as e-paper such as electrochromic display (Schoot et al., 1973; Kobayashi et al., 2008), electrowetting display(Beni and Hackwood, 1981), MEMS display(Miles et al., 2003; Taii et al., 2006), liquid powder display(Hattori et al., 2004; Sakurai et al., 2007) and cholesteric liquid crystal display (Berreman and Heffner, 1980; Tamaoki, 2001). Twisting-ball display(Howard et al., 1998; Sheridon et al., 1999) is one of them. Twisting-ball display was initially invented in the 1970s by Dr. Sheridon. The structure of a twisting-ball display is shown in Figure 1. “Janus ” particles which have black and white hemisphere are sandwiched between two electrodes. Each hemisphere of the particles is differently charged and the direction of the particles can be controlled by applied voltage. Displaying grayscale images on twisting-ball display is difficult, but a quasi-grayscale image can be displayed by “dithering” which is a common method to obtain quasi-grayscale images on a monochrome display. Janus particles are suspended by silicone-oil-filled cavities in silicone rubber sheet so that they do not move around and aggregate. The size of Janus particles influences various performance of the display such as resolution, contrast and driving voltage. With small particles, the resolution is improved and driving voltage is reduced, but contrast is spoiled because a small amount of light penetrate front-side hemisphere and color of back-side hemisphere is seen. Considering their balances, typical size

*E-mail address: komazaki@dt.k.u-tokyo.ac.jp (Corresponding author).

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 331 of 404

of Janus particles is around 100 μm with resulting low resolution (254 ppi at a theoretical maximum). However, large-sized twisting-ball display is easily fabricated due to its simple structure and fabricating process. Therefore, large-sized sign board or digital signage will be a suitable application of this display. Commercial products are very little at present, but an outdoor timer is on sale from molten, Japan (Figure 2).

Although writing and drawing are an important function of paper, studies on writing function of e-"paper" are limited. There are some e-reader products with writing functions, but these functions are achieved by the same technology as smart phones using touch sensors. For these devices, latency is inevitable while writing or drawing because refresh rate of e-paper is generally low and processing time is needed. To eliminate latency and realize smooth writing, direct driving of pixels by writing stimuli such as writing pressure and magnetic field by the magnetic pen is effective (Figure 3). For this direct drive configuration, product cost can be reduced because touch sensors and processors are not required. For example, "Boogie Board" (Kent Displays, US) achieved very smooth writing by using pressure-sensitive cholesteric liquid crystal display without touch sensors(Lee et al., 2007; Montbach, 2008). Figure 4 shows a comparison of handwriting on an e-reader (PRS-350, SONY) and Boogie Board. The latency of 0.2-0.5 sec was observed on the e-reader while Boogie Board had no latency. In a similar concept, we had developed a handwriting-enabled twisting-ball display by just adding magnetic nanoparticles into one hemisphere of Janus particles(Komazaki and Torii, 2013, 2014; Komazaki et al., 2015). The structure of this display is shown in Figure 5. The color of the display (black/white) can be controlled by applied voltage like conventional twisting-ball display and handwriting in the absence of voltage by using a magnetic pen is also possible. Because magnetic nanoparticles have superparamagnetism, with which they have no remanent magnetization, there is no magnetic interaction between each particle when the magnetic pen is not applied so that particles maintain their direction(Ghosh et al., 2008). Therefore, this display is bistable not only for electric control, but also magnetic handwriting. We describe the fabrication method, performance and application of this display below.

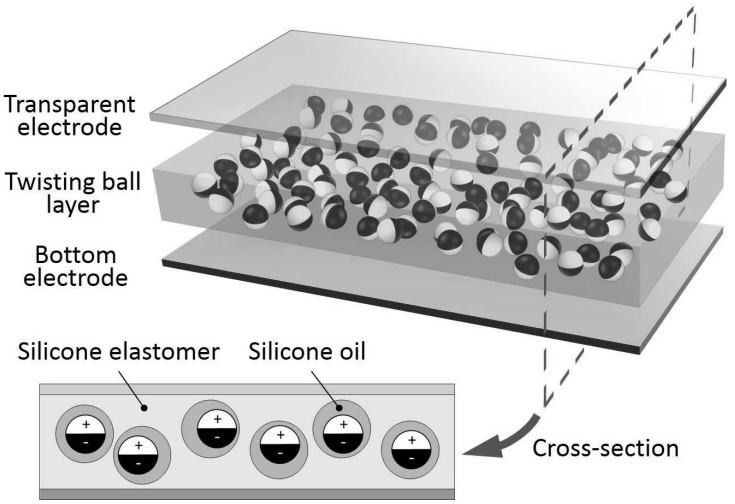


Figure 1. Structure of a twisting-ball display.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 332 of 404



Figure 2. Outdoor timer utilizing twisting-ball display (blue/white) which achieves high contrast under strong sunlight.

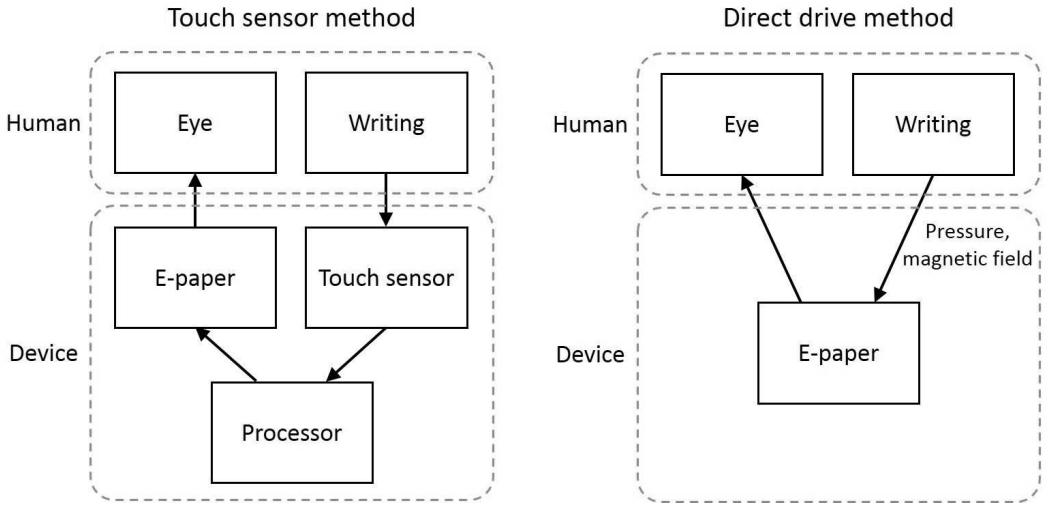


Figure 3. Difference between handwriting system of touch sensor method and direct drive method.

2. Materials and Methods

Janus particles for handwriting-enabled twisting-ball display were synthesized by a microfluidic method(Nisisako et al., 2006). Setup for microfluidic Janus particle synthesis is shown in Figure 6. In this method, Janus droplets were formed in the microchannel from black and white monomers (Figure 7) and formed precursor droplets were polymerized in polymerization bath by thermal polymerization. To enable handwriting, 2 wt% magnetic nanoparticles were added into black monomer. To fabricate a twisting-ball display, silicone rubber sheet in which Janus particles were densely embedded was formed and soaked in silicone oil. Through soaking, the sheet absorbed silicone oil and swelled, by which oil-filled

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 333 of 404

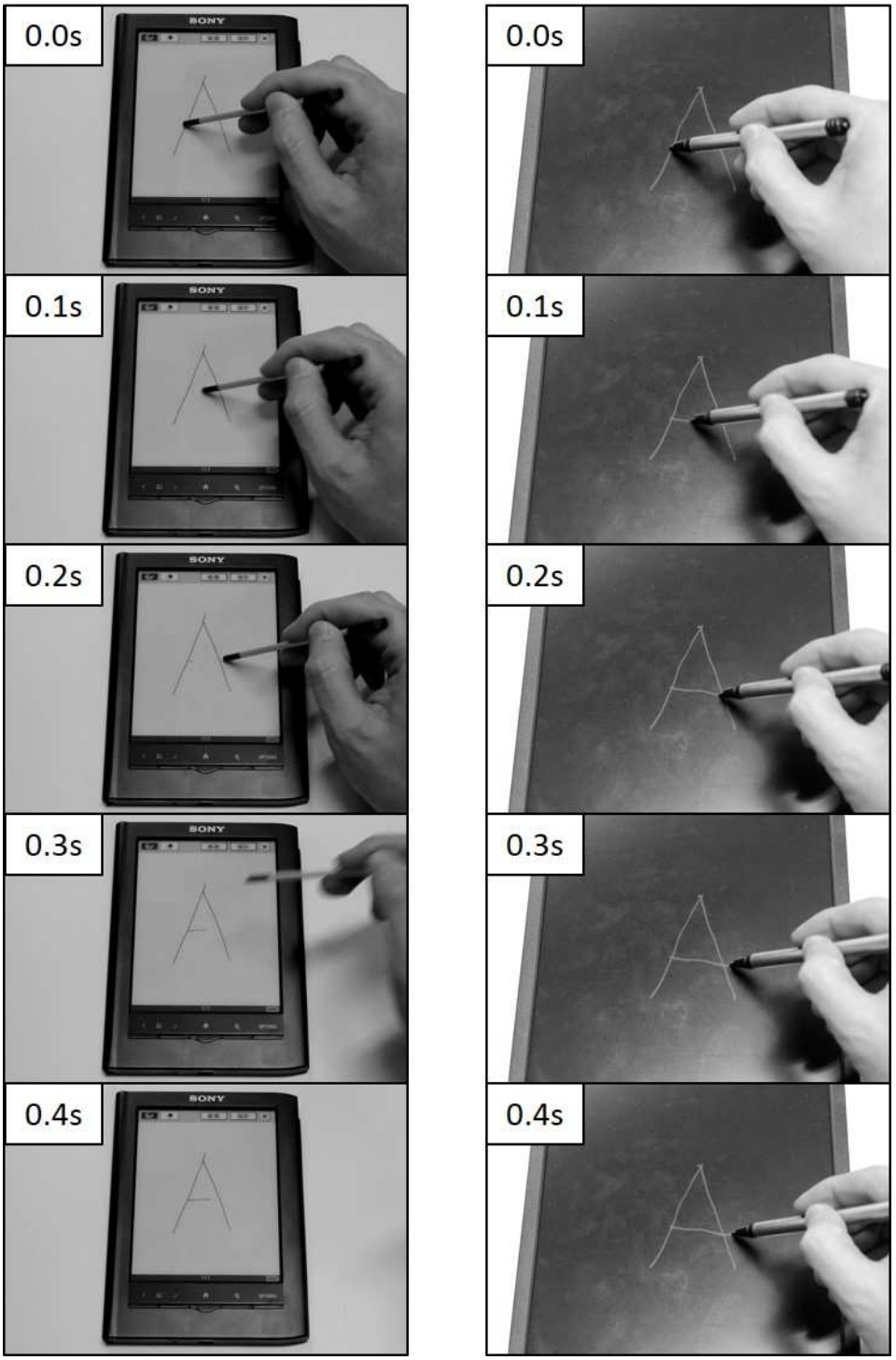


Figure 4. Comparison of handwriting on an e-reader (PRS-350, SONY) and Boogie Board. Latency of 0.2-0.5 sec was observed on the e-reader.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 334 of 404

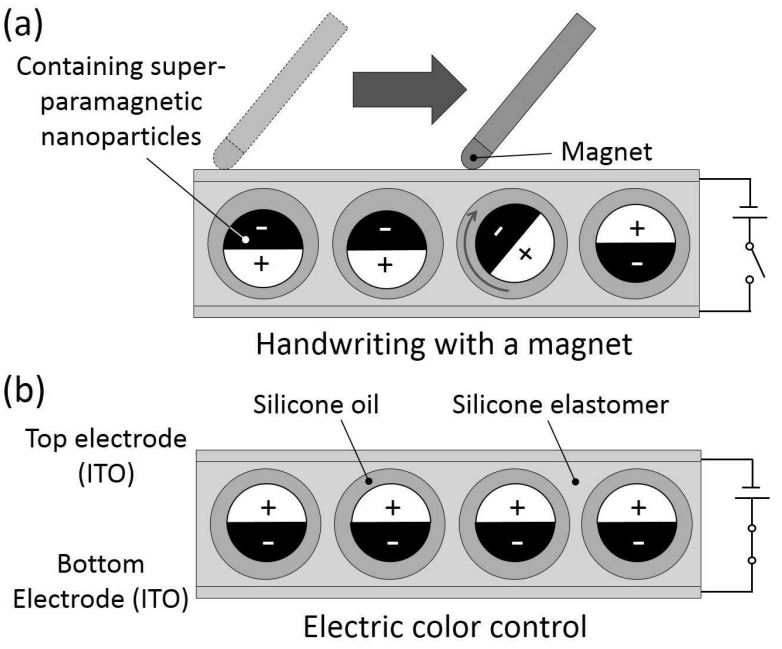


Figure 5. Structure of handwriting-enabled twisting-ball display(Komazaki et al., 2015). (a) Handwriting with a magnet on the display. (b) Electric color control of the display.

cavities for particle rotation were formed. The sheet was sandwiched by two ITO coated glass plates and twisting-ball display was fabricated.

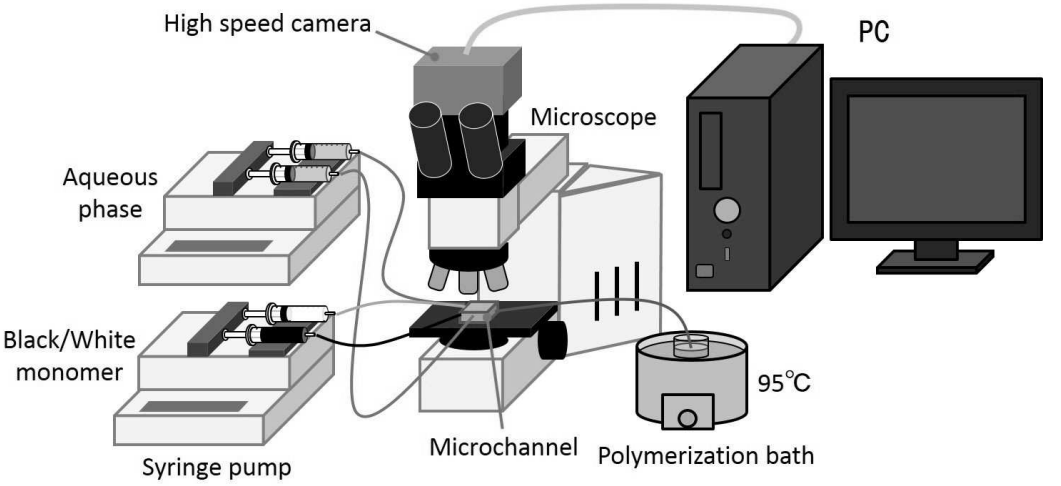


Figure 6. Setup for microfluidic Janus particle synthesis(Komazaki and Torii, 2013).

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 335 of 404

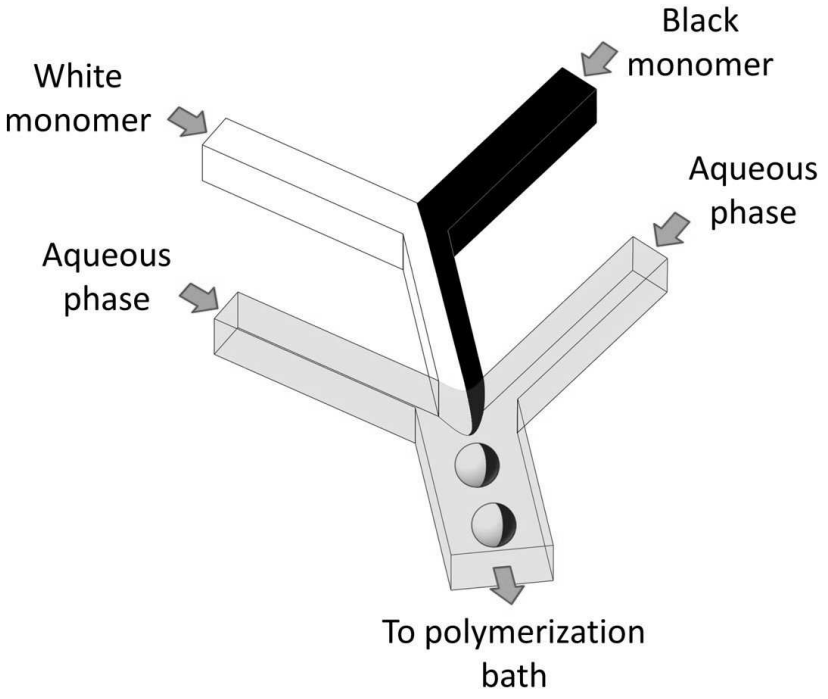


Figure 7. Schematic illustration of Janus droplet formation in a glass microchannel(Komazaki et al., 2015).

3. Results and Discussion

Synthesized Janus particles are shown in Figure 8. Mean diameter of the particles was 142 μm and coefficient of variation (CV) was 2.4% which represented high monodispersity. Although black and white monomer mixed a little and geometry of each hemisphere was not perfect, the particles still had black and white region on their surfaces. The thickness of the silicone rubber sheet after swelling was 0.45 mm. By applying $\pm$80 V, the color of the display could be controlled as shown in Figure 9. Handwriting with a small magnet in absence of voltage was also possible with no latency (Figure 10). Though contrast was not so good, it will be improved by inhibiting mixing of black and white monomer and increasing the concentration of pigments. By changing the size of magnet, line width could be changed. Written characters were maintained as long as voltage was not applied. This means magnetic nanoparticles maintained their superparamagnetism and there was no magnetic interaction among particles by remanent magnetization. Although no latency was observed and handwriting on the display was quite smooth, handwriting right after the voltage was turned off was impossible (Figure 11). The recovery time of handwriting was measured as 5 sec. This phenomenon was attributed to the capacitance of the display. Because silicone rubber, silicone oil and Janus particles were insulants, charges on top and bottom electrodes were maintained for a certain time and they prevented Janus particles from flipping. To eliminate this phenomenon, increasing the number of magnetic nanopar-

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 336 of 404

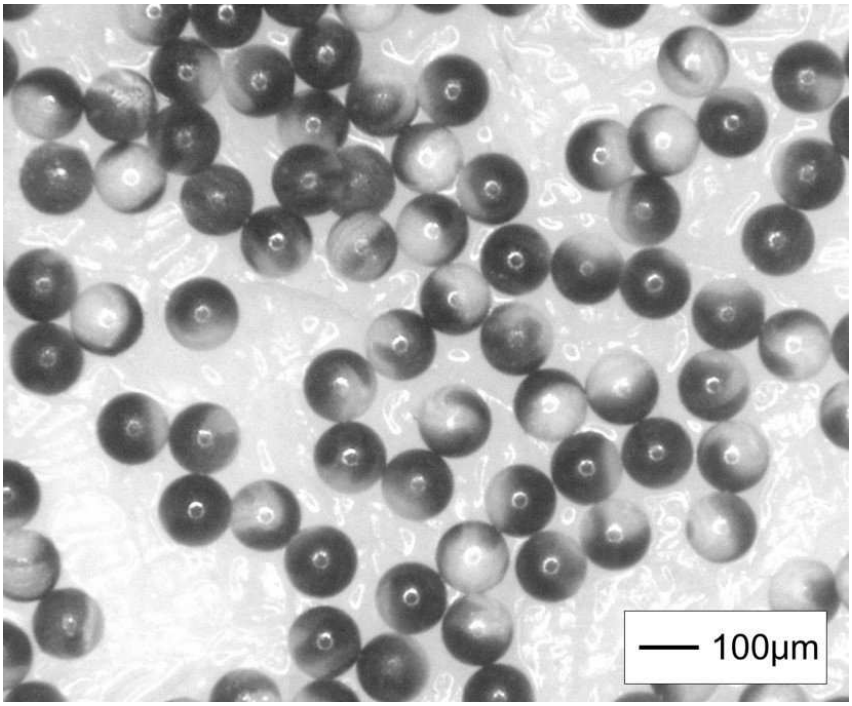


Figure 8. Janus particles containing magnetic nanoparticles in black hemisphere(Komazaki et al., 2015). Mean diameter was 142 μ and CV was 2.4%.

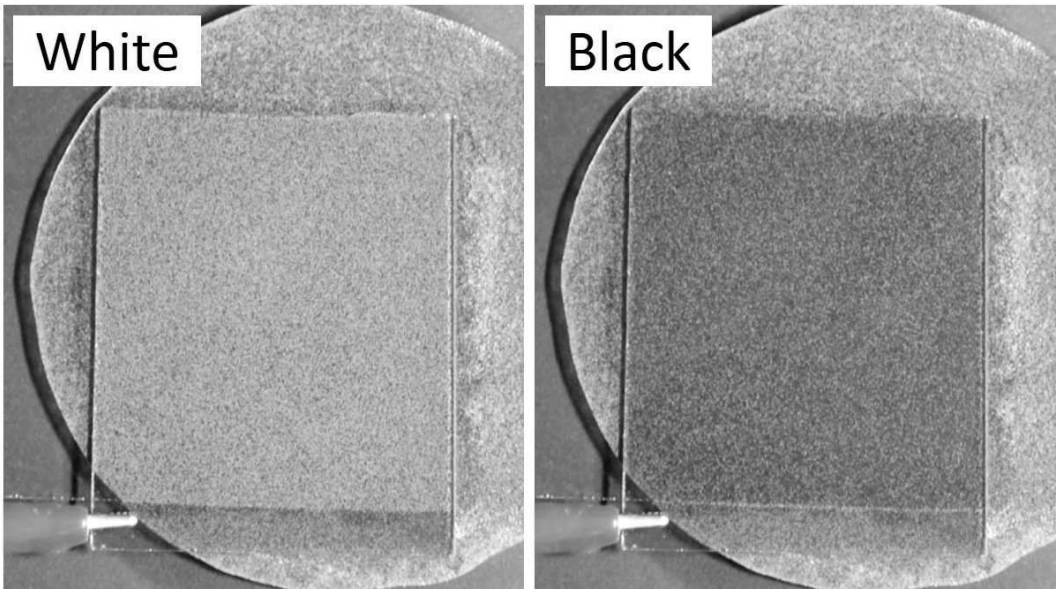


Figure 9. Color control of the display by applying $\pm$80 V(Komazaki et al., 2015).

ticles was effective. For Janus particles containing 4 wt% magnetic nanoparticles in black hemisphere, handwriting even while 80 V was applied was possible (Figure 12). Therefore,

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 337 of 404

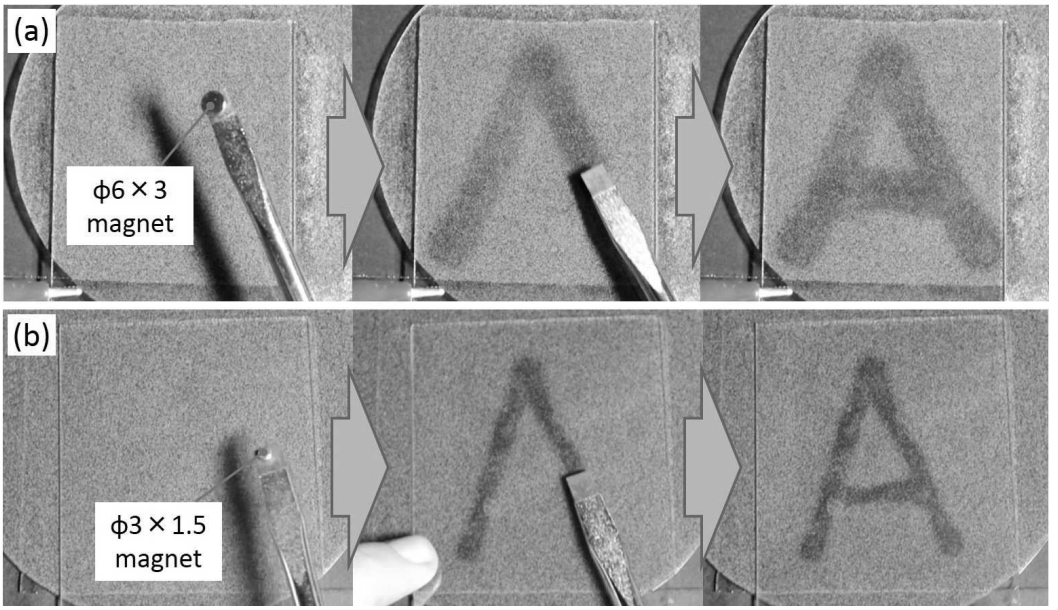


Figure 10. Handwriting with a small magnet without voltage. (a) with a ϕ6 x 3 mm neodymium magnet. (b) with a ϕ3 x 1.5 mmneodymium magnet(Komazaki et al., 2015).

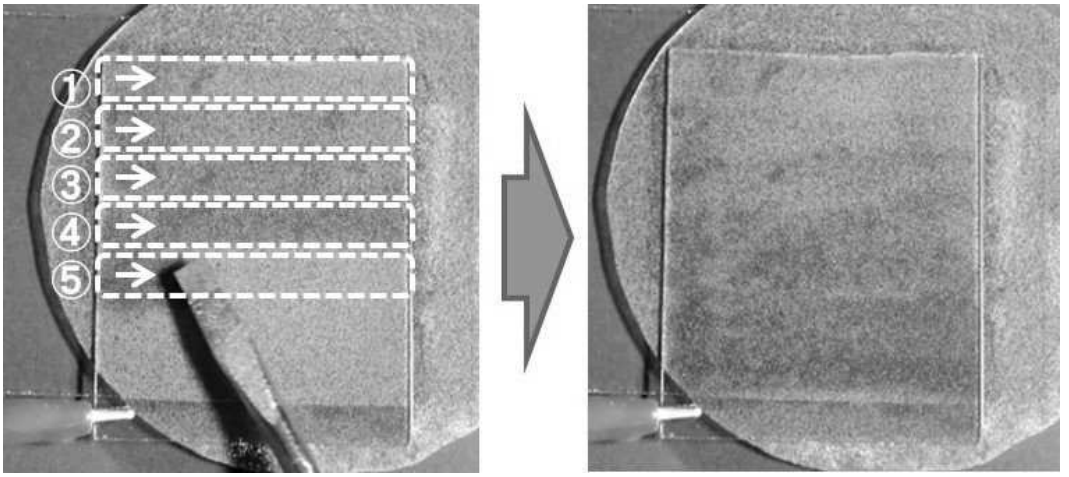


Figure 11. Impossibility of handwriting right after voltage is applied(Komazaki and Torii, 2014). Handwriting was started when the voltage was turned off. The upper region of the display remained white.

a certain amount of magnetic nanoparticles is needed to achieve anytime handwriting. Another way to avoid this phenomenon which can save the number of magnetic nanoparticles is discharging the charges on the electrodes thorough a resistor immediately after voltage is removed (Figure 13). Although this display itself does not accept input, written character sensing will be easily achieved with sensors because inverted image is shown on the back side of the display if the bottom electrode is transparent (Figure 14-a). For example, the

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 338 of 404

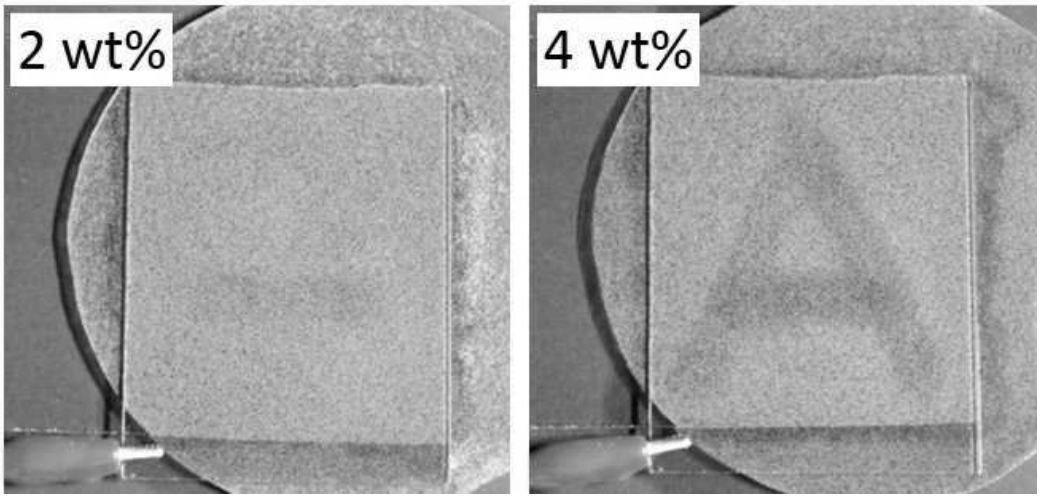


Figure 12. Handwriting while 80 V is applied to the display(Komazaki et al., 2015). For Janus particles containing 4 wt% magnetic nanoparticles in black hemisphere, handwriting even while 80 V was applied was possible

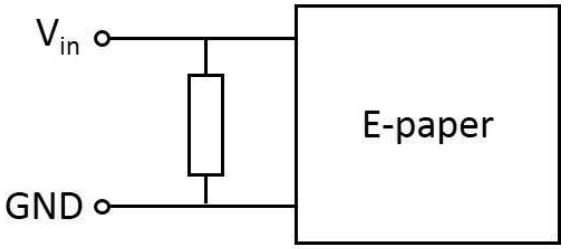


Figure 13. Connection of a resistor to discharge the charges on the electrodes immediately after voltage is removed.

optical scanner can be placed on the back side of the display (Figure 14-b). In this case, only offline handwriting recognition is possible. To achieve online recognition, optical sensor matrix is required (Figure 14-c). In addition, if the detection of slight current on the electrodes induced by rotation of polarized Janus particles is possible, handwritten strokes can be acquired even without optical sensors. By these methods, large handwriting input devices will be produced at low cost.

Conclusion

In this chapter, we discussed background, structure, fabrication method, performance and applications of handwriting-enabled twisting-ball display. The display achieved smooth handwriting in addition to conventional electronic display function by quite simple method. Although contrast is not so good at this moment, it will be improved by inhibiting mixing of black and white monomer and increasing the concentration of pigments. As for applications of this display, small mobile devices will not be suitable due to low resolution. However,

Cerence Operating Company, Ex. 1011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 339 of 404

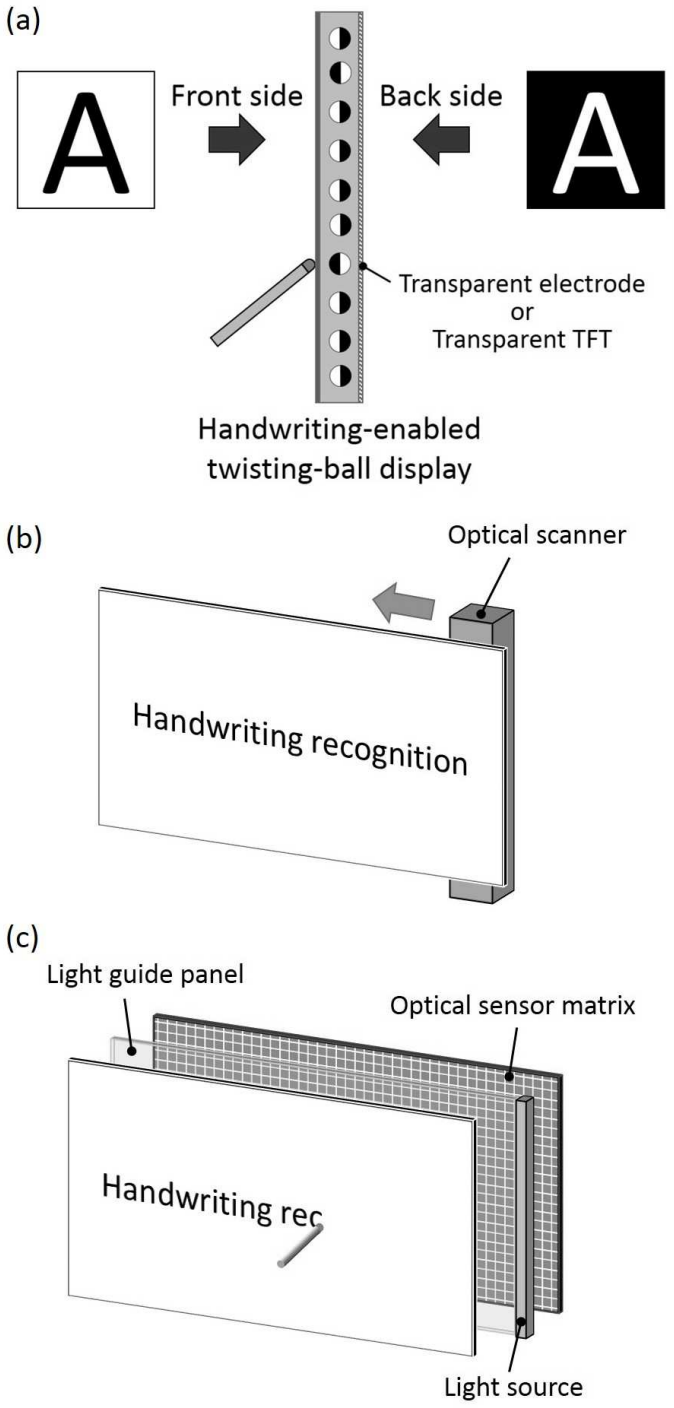


Figure 14. Written character sensing from back side image. (a) Images on front and back side of the handwriting-enabled twisting-ball display. (b) Offline handwriting recognition with optical scanner. (c) Online handwriting recognition with optical sensor matrix.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 340 of 404

a large-size display is easily fabricated because of simple structure. Therefore, large-size devices such as electronic whiteboards will be a suitable application. Although multicolor and movie are impossible, much cheaper and larger electronic white board systems than conventional systems with expensive and fragile large LCDs or projectors will be realized by using this handwriting-enabled twisting-ball display. Cheaper electronic whiteboards will contribute to digitalization of office and education, especially in developing countries.

References

Beni, G. and Hackwood, S. (1981). Electro-wetting displays. *Applied Physics Letters*, 38(4):207–209.

Berreman, D. W. and Heffner, W. R. (1980). New bistable cholesteric liquid-crystal display. *Applied Physics Letters*, 37(1):109–111.

Ghosh, A., Sheridon, N. K., and Fischer, P. (2008). Voltage-controllable magnetic composite based on multifunctional polyethylene microparticles. *Small*, 4(11):1956–1958.

Hattori, R., Yamada, S., Masuda, Y., Nihei, N., and Sakurai, R. (2004). A quick-response liquid-powder display (QR-LPD®) with plastic substrate. *Journal of the Society for Information Display*, 12(4):405.

Howard, M. E., Richley, E. A., Sprague, R., and Sheridon, N. K. (1998). Gyricon electric paper. *Journal of the Society for Information Display*, 6(4):215.

Jacobson, J., Comiskey, B., Albert, J. D., and Yoshizawa, H. (1998). *Nature*, 394(6690):253–255.

Kobayashi, N., Miura, S., Nishimura, M., and Urano, H. (2008). Organic electrochromism for a new color electronic paper. *Solar Energy Materials and Solar Cells*, 92(2):136–139.

Komazaki, Y., Hirama, H., and Torii, T. (2015). Electrically and magnetically dual-driven janus particles for handwriting-enabled electronic paper. *Journal of Applied Physics*, 117(15):154506.

Komazaki, Y. and Torii, T. (2013). Writable electronic paper based on twisting ball type electronic paper. In *The 20th International Display Workshops*, pages 1340–1343.

Komazaki, Y. and Torii, T. (2014). Power-saving bar indicator for applied voltage utilizing twisting ball technology. In *The 20th International Display Workshops*, pages 1192–1193.

Lee, D. W., Shiu, J. W., Sha, Y. A., and Chang, Y. P. (2007). 6.3: Writable cholesteric liquid crystal display and the algorithm used to detect its image. *SID Symposium Digest of Technical Papers*, 38(1):61–64.

Miles, M., Larson, E., Chui, C., Kothari, M., Gally, B., and Batey, J. (2003). Digital paper™ for reflective displays. *Journal of the Society for Information Display*, 11(1):209.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 341 of 404

Montbach, E. (2008). Flexible electronic flat-panel displays find novel uses. *SPIE Newsroom*.

Nisisako, T., Torii, T., Takahashi, T., and Takizawa, Y. (2006). Synthesis of monodisperse bicolored janus particles with electrical anisotropy using a microfluidic co-flow system. *Advanced Materials*, 18(9):1152–1156.

Ota, I., Ohnishi, J., and Yoshiyama, M. (1973). Electrophoretic image display (epid) panel. *Proceedings of the IEEE*, 61(7):832–836.

Sakurai, R., Ohno, S., ichi Kita, S., Masuda, Y., and Hattori, R. (2007). Color displays and flexible displays using quick-response liquid-powder technology for electronic paper. *Journal of the Society for Information Display*, 15(2):127.

Schoot, C. J., Ponjee, J. J., van Dam, H. T., van Doorn, R. A., and Bolwijn, P. T. (1973). New electrochromic memory display. *Applied Physics Letters*, 23(2):64–65.

Sheridon, N. K., Richley, E. A., Mikkelsen, J. C., Tsuda, D., Crowley, J. M., Oraha, K. A., Howard, M. E., Rodkin, M. A., Swidler, R., and Sprague, R. (1999). The gyricon rotating ball display. *Journal of the Society for Information Display*, 7(2):141.

Taii, Y., Higo, A., Fujita, H., and Toshiyoshi, H. (2006). A transparent sheet display by plastic MEMS. *Journal of the Society for Information Display*, 14(8):735.

Tamaoki, N. (2001). Cholesteric liquid crystals for color information technology. *Advanced Materials*, 13(15):1135–1147.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 342 of 404

In: Handwriting: Recognition, Development and Analysis ISBN: 978-1-53611-937-4
Editors: Byron L. D. Bezerra et al. © 2017 Nova Science Publishers, Inc.

Chapter 13

SPEED AND LEGIBILITY: BRAZILIAN STUDENTS PERFORMANCE IN A THEMATIC WRITING TASK

Monique Herrera Cardoso* *and Simone Aparecida Capellin*†
Investigation Learning Disabilities Lab at the Speech
and Hearing Sciences Department,
São Paulo State University "Júlio de Mesquita Filho" - UNESP,
São Paulo, Brazil

1. Introduction

The domain of writing skills still constitutes an important goal which should be achieved by school children (McCarney et al., 2013), and therefore deserves greater attention from educators and health professionals (Feder and Majnemer, 2007), because it plays an important role in school performance, with relevant implications for the motor and cognitive development (Accardo et al., 2013).

When it comes to academic performance, it can be observed that the academic skills of the students from elementary to high school, are measured by tests involving proficient writing, either by bimonthly tests or even by the Secondary Education National Examination (Enem). However, a poor handwriting may mask the academic ability of a student, because teachers tend to give higher grades to legible handwriting tasks than to works illegible handwriting tasks (McCarney et al., 2013).

This negative perception tends to be come highlighted as schooling increases especially when handwriting is confusing and difficult to read (Piek et al., 2008). When the child is competent in other areas, the difficult issues concerning writing can be attributed to laziness or lack of motivation, rather than a specific learning difficulty that affects language production in its written form (Berninger et al., 2009).

According to international studies, handwriting ability is responsible for alteration in quality of texts of students, from the first year of primary school (Graham et al., 1997) to undergraduate students (Peverly, 2006). The researches (Medwell and Wray, 2008) reported

*E-mail address: moniquehc@gmail.com.
†E-mail address: sacap@uol.com.br.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 343 of 404

in their study that cognitive processes involved in writing compete with other processes, such as planning and generating ideas. Thus, when the child has not reached a writing automation level, he/she starts to use working memory features to focus attention for forming letters and words, generating a negative effect upon his/her textual composition quality.

This finding was replicated in a study comprising adult writers (Tucha et al., 2008), whose fluency was proven impaired when they were required to pay close attention to their writing. While the child still has to consciously focus on the writing mechanism, the amount of working memory available to concentrate on the content in which they are writing is reduced (McCarney et al., 2013; Prunty et al., 2013).

However, in Brazil, there are two problems: (1) the calligraphy teaching methodology is not standardized, i.e., each school emphasizes and defines how much the student should practice. (2) Specific assessment instruments, based on appropriate criteria for school children, which measure the performance speed and observation of writing legibility aspects, are not available.

The Brazilian researches (Cardoso et al., 2014) made the translation and adaptation of the Detailed Assessment of Speed of Handwriting - DASH (Barnett et al., 2007) for the Brazilian population. DASH is a standardized procedure with a representative stratified sample of the UK and identifies children from 09 to 16 years, with handwriting difficulties.

It comprises five tasks: four writing tasks (Task 1 - Copy Best; Task 2 - Alphabet Writing, Task 3 - Copy Fast, Task 5 - Free Writing), and one of them is a perceptual-motor skill measure, i.e., it does not involve language linguistic aspects (Task 4 - Graph Speed).

It is clear that the thematic Free Writing Task is more similar to an exam / test environment, for the student Therefore, acknowledging the expected from profficient writers, as legibility and writing speed, it becomes possible to help professionals working in educational area to identify students presenting difficulties in writing skills development and to quantify the underformance as for the chronological age and school grade.

Given the above, this chapter presents the performance of the Brazilian students during the Thematic Free Writing, according to DASH Brazilian version.

2. Method

This study was approved to the Ethics Committee of the Faculty of Philosophy and Sciences, São Paulo State University “Júlio de Mesquita Filho”, UNESP - Marília- São Paulo, Brazil, under protocol number 0444/2012.

This study comprised 05 public schools, located in the city of Marília (SP) and Region: one state school and one municipal school, located in a rural area; one state and one municipal school, located in the south area; and one municipal school located in the north area. These schools take part of trainings conducted by the research group of the Research Laboratory of Learning Deviations - LIDA - Department of Speech Pathology, UNESP Marília (SP), and therefore, the students are often accompanied by the entire staff (pedagogues, speech language therapists, occupational therapists, neuropsychologists, and child neurology) thus, they are diagnosed and referred for intervention, when necessary.

The age of the students selected to participate in this study ranged from 09 to 16 years and 11 months, and the Informed Consent was signed by parents or guardians, and they

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 344 of 404

should not present constant notes of sensory, motor or cognitive impairments and / or hearing visual or motor complaints in school records. Failure to comply with at least one of the above criteria, automatically excluded the student.

As a selection criterion, a cognitive assessment test was employed (Raven Progressive Matrices), for excluding cases of mental retardation, and also a survey was conducted along with the pedagogical coordinators of the students who had school complaints, psych affective problems or other diagnoses (e.g., autism, attention deficit and hyperactivity disorder, dyslexia and others), and were excluded from the sample data for this study.

According to the criteria, 658 students took part in this study, both genders, from 09 to 16 years and 11 months old, and were divided into age groups (Table 1).

Table 1. Distribution of the number of school children evaluated according to the study groups, age and gender

Groups	Ages	N	Males		Females	
			N	(%)	N	(%)
GI	09 to 09 years and 11 months	102	58	57	44	43
GII	10 to 10 years and 11 months	99	44	44	55	56
GIII	11 to 11 years and 11 months	86	41	47	45	53
GIV	12 to 12 years and 11 months	90	43	48	47	52
GV	13 to 13 years and 11 months	87	39	45	48	55
GVI	14 to 14 years and 11 months	68	28	41	40	59
GVII	15 to 15 years and 11 months	61	27	44	34	56
GVIII	16 to 16 years and 11 months	65	34	52	31	48
Total		658	314	47,65	344	52,35

Data collection was held in groups comprising 15 to 20 students in one unique session, not exceeding 50 minutes. The students should use lined sheets, their own pencils and were allowed to use erasers.

To perform Task 5 (Thematic Free Writing), the student was requested to write an essay on “My Life” for 10 minutes; however, every two minutes the student should make a mark in the text, which allowed us to monitor the production rate of the student, in different periods of time.

For educational and clinical purposes, these data can be very informative. For example, they allow us to distinguish from the child which is always slow throughout the 10-minute period; and the child that writes a lot for a minute, and then simply runs out of ideas. There is no doubt that a task like this is very similar to an examination environment, for the student.

To calculate the student writing speed, taking into account the writing legibility, it was initially necessary to identify the legible words (step 1). After identification, it was possible to calculate the writing speed (step 2).

Step 1 - Definition of the readable words

In order to obtain reliable and trustworthy results, the writing samples of the students were judged as for legibility, by four professionals who work with calligraphy and analysis of school spelling: a speech therapist, two occupational therapists and a pedagogue, i.e.

Cerence Operating Company, Ex. 2017
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 345 of 404

professionals who have the same concept concerning calligraphy, but work in different areas. As precaution measure four judges were selected, in case any of them could not perform the analysis within the time proposed by the researcher.

A meeting was held with the judges and the researcher, and it was handed a script guidance, explaining how the samples should be analyzed; a 2-hour training and, finally, clarifications and doubts were solved. At the end of the meeting, the researcher provided her contacts (e-mail and phone) if, during the analysis of the samples, the judges had any questions and / or difficulty.

The writing samples were scanned, and the names of the students were ommitted and delivered to each judge in a pendrive containing 08 files, Microsoft Office PowerPoint Program, version 97-2003 or updated. Each file was composed by samples concerning a specific age, however, only the researcher knew the age which corresponded to a particular file. This measure was to ensure that the judges did not know the age of the student which corresponded to that writing sample, and consequently did not make comparisons and judgments related to the age of the student.

Judges should read every word written by the student, only once. If they did not understand, they should not insist on re-reading, or even “try” to understand, making use of the phrase context. Each word read by the judge should be classified as:

(A1) Legible: word which the judge easily decoded, regardless the context of the sentence. Sometimes, there were poorly formed letters in a word, which, if considered out of the word, would not be legible. However, if the entire word could be still legible, it should be classified as such. When the judge classified the word as legible, no mark should be performed.

(A2) Partially legible: word which the judge was able to read, however, had difficulty to decode it. When the judge described the word as partially legible, he made a blue rectangle around the word.

(A3) Totally illegible: word which the judge failed to read, due to difficulty in decoding it. When the judge described the word as completely illegible, he made a red rectangle around the word.

Cronbach’s alpha statistic test was employed (or, simply best known as Cronbach test) to check the level of reliability in terms of so-called ’internal consistency’ of the observed values. This, in turn, presented a variation between 0.700 to 1.000, for all variables analyzed by the judges, which allowed us to consider a sample with ’high’ reliable level, which provides an unbiased sample to this task.

After this stage, it can be inferred that the concept of legibility is not divergent among the professionals, and therefore, the analysis of the pedagogue judge was randomly selected to be employed for speed calculation (step 2).

Step 2 - Calculation of the writing speed

The writing speed calculation was performed according to the performance presented by each study group, in the thematic writing task. It was calculated by measuring the readable written words per minute (wpm), that is, the number of words written by each group, except for the number of words considered unreadable.

For the Thematic Writing Task, the speed was calculated every two minutes, and also during the activity as a whole (10 minutes).

From the analyses of the samples, data were tabulated in Microsoft Office Excell spreadsheet, version 2010, then, descriptive analysis and statistical data were conducted.

Cerence Operating Company, Ex. 2011
Signifying Descriptive Analysis and Statistical Data, Vol. 1, p. 01267
Complimentary Contributor Copy
Page 346 of 404

3. Results and Discussion

Age constitutes a determining factor for the amount of legible words, so, it was observed that younger students underperformed (in a 10-minute task), when compared to older students, that is, writing fluency differed for each age group (Table 2).

However, age is not a determining factor for the calligraphic quality, since there was no statistically significant difference, when compared the performance of groups simultaneously referring to variables “partially legible” and “illegible” (Table 2). These results corroborate the literature (Overvelde and Hulstijn, 2011), reporting that students from nine years of age have an automatic and organized writing, because they have acquired the mechanical sense in writing, that is, they know how to hold a pencil, to form isolated letters and put the letters smoothly and fluently (McCarney et al., 2013), which in turn, enables them to spell legible words.

Table 2. Legibility according to age group

Legible	**Group**	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	< 0,001*
	N	102	99	86	90	87	68	61	65	
	Mean	78,93	92,96	94,14	105,14	111,66	106,16	96,18	111,2	
	SD	27,8	34,71	29,84	39,29	37,02	41,18	41,91	41,92	
Parcially legible	**Group**	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	0,838
	N	102	99	86	90	87	68	61	65	
	Mean	1,89	2,36	3,03	3,26	1,97	2,31	2,1	2,4	
	SD	2,02	3,71	4,25	4,84	2,3	2,91	3,06	2,93	
Illegible	**Group**	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	0,366
	N	102	99	86	90	87	68	61	65	
	Mean	0,3	1,09	1,02	1,19	0,8	0,72	0,51	0,83	
	SD	0,74	5,07	4,13	3,24	1,89	1,74	1,31	2,45	

Taking into account only the legible words, the writing speed was calculated for each studied group (GI, GII, GIII, GIV, GV, GVI, GVII, GVIII), every two minutes of the task, and throughout the 10 minutes (Table 3).

With these results it can be observed that students tend to write faster (words per minute) during the first two minutes of the task, except the students of 14, 15 and 16 (GVI, GVII and VIII), who presented higher speed from the second to fourth minute of the task. These findings show that students easily idealize the beginning of their textual composition, releasing most of the attention resources and working memory to form letters and words (Graham et al., 2006). However, with the task exposure time, the student must focus on the content that he/she is writing, thus decreasing the speed for spelling the sequence of letters, words, and finally, the construction of phrases.

It becomes possible to observe that writing speed decreases as task exposure time increases, except for 12 years-old-students (GIV), who increase their speed in the variable “4 to 6 minutes”. This writing speed reduction can be justified by tiredness / fatigue of the students during the tasks development. This finding supports the study (O’Mahony et al., 2008), reporting that writing speed is impacted by several cognitive, motivational and physiological factors, such as muscle fatigue, due to a prolonged writing period.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 347 of 404

Table 3. Writing speed, according to age, for thematic Free Writing Task

	Group	I	II	III	IV	V	VI	VII	VIII	Sig (p)
Up to 2 min	N	102	99	86	90	87	68	61	65	< 0,001*
	wpm	19,57	22,53	24,13	24,17	26,78	26,74	21,92	27,66	
	SD	7,4	9,73	8,6	9,68	10,28	11,64	12,18	11,27	
	Grupo	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	
From 2 to 4 min	N	102	99	86	90	87	68	61	65	< 0,001*
	wpm	16,73	19,23	23	21,99	25,79	27,25	22,48	28,48	
	SD	6,62	8,06	9,26	8,65	8,04	13,49	9,67	10,21	
	Grupo	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	
From 4 to 6 min	N	102	99	86	90	87	68	61	65	< 0,001*
	wpm	16,2	18,53	20,2	22,34	24,85	21,96	19,82	24,14	
	SD	8	8,2	9,18	10,04	9,6	10,51	13,18	11,1	
	Grupo	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	
From 6 to 8 min	N	102	99	86	90	87	68	61	65	0,017
	wpm	15,16	17,11	16,37	20,9	18,98	18,29	16,87	18,38	
	SD	7,36	8,98	10,83	10,39	12,01	13,83	11,86	12,26	
	Grupo	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	
From 8 to 10 min	N	102	99	86	90	87	68	61	65	0,816
	wpm	11,28	15,57	10,44	15,74	15,25	11,93	15,1	12,54	
	SD	8,89	11,28	10,63	11,92	13,11	12,7	13,32	13,19	
	Grupo	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	
During 10 min	N	102	99	86	90	87	68	61	65	< 0,001*
	wpm	78,93	92,96	94,14	105,14	111,66	106,16	96,18	111,2	
	SD	27,8	34,71	29,84	39,29	37,02	41,18	41,91	41,92	

Subtitles: GI - 9 years; GII - 10 years; GIII - 11 years, GIV - 12; GV - 13 years; GVI - 14 years; GVII - 15; GVIII - 16 years. wpm - words per minute; SD - standard deviation; Jonckheere-Terpstra test. (P) = 0.005

Then, the Jonckheere-Terpstra test was applied, in order to verify possible differences among the eight groups, when compared simultaneously, and there were statistically significant differences in variables “up to 2 minutes”, “from 2 to 4 minutes ”, “from 4 to 6 minutes”, and “during the 10 minutes”. That is, there was no difference in writing speed, from the 6th minute of the task, when groups were compared.

This finding questions if the task should really take 10 minutes, considering that fatigue may have been a detrimental factor for the students’performance of the eight groups. Another factor for justifying this finding would be related to the pauses of students, while writing, that is, a time variable (Benbow, 1995), as it should be necessary to investigate, if during the 04 final minutes of the task, the students present more breaks, either for fatigue reasons, or due to difficulty in continuing the suggested theme for writing.

According to the literature (Olive, 2010; Sumner et al., 2013), writing was considered as a continuous movement, interrupted by ’breaks’, i.e., temporary interruptions in the written trace. These breaks are “normal”, whereas they are simply imposed by the text to be written, as in the case of spaces between words and between letters (Paz-Villagrán et al., 2014). However, studies have shown that dyslexic students take longer to write the same text, not because they have a slower movement, but due to the fact that they make more breaks, or longer interruptions when compared with proficient writers (Rosenblum et al., 2003a,b).

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 348 of 404

In addition to investigation concerning breaks, one should also think about the ergonomic factors such as posture during writing and / or pencil gripping, as they may change along this task (Rosenblum et al., 2006; Tomchek and Schneck, 2006), and because they are factors which can lead to difficulties for proper writing performance, hence causing reduction in legibility, pain and fatigue in the upper limbs (de Almeida et al., 2013; Sassoon, 2004), leading to decreased writing speed.

When analyzing Figure 1, it can be observed from 9 to 13 years, increased writing speed; however, from 14 to 16 years the speed has declined, with little variation among them.

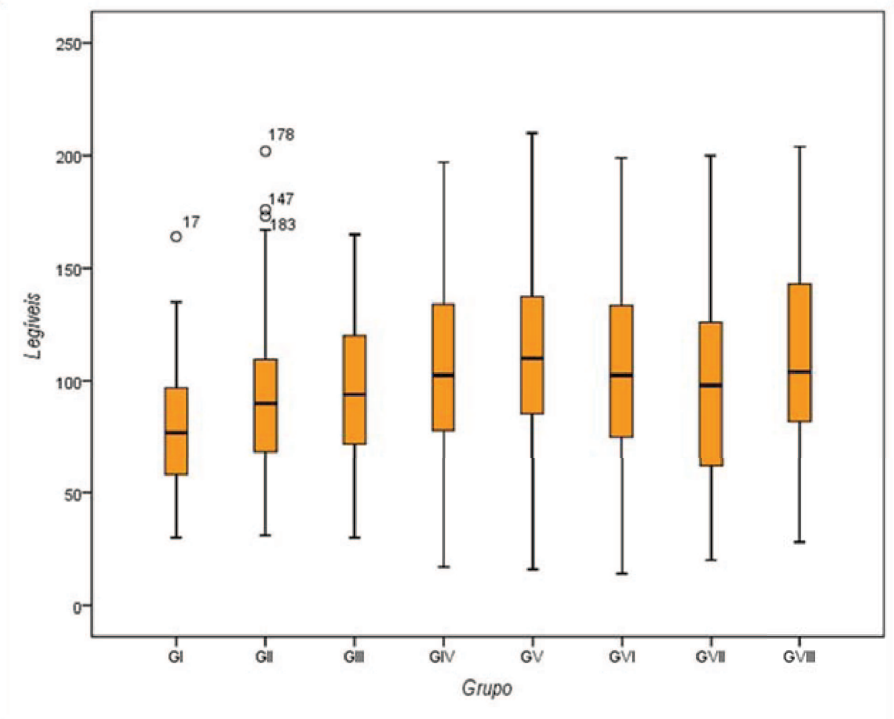


Figure 1. Box plot for writing speed score in task 05, per group. The box represents 50% of the results (and the bottom line, percentile 25, and the top line, percentile 75), the line inside the Box represents the mean found, and the external lines of the box represent the maximum and minimum values found

This increase in speed due to the increasing age can be justified, because in early learning, movements are slow and guided by visual and kinesthetic feedback (Chartrel and Vinter, 2006). That is, during handwriting, information, such as the pressure on pencil and paper, the positioning of the fingers and hand, the direction of pencil movements and mistakes, are stored in memory, to be recalled when writing is repeated (Almeida, 2013). With practice, writing becomes automatic and control of writing coordinated movements are improved with age and education, favoring increased writing speed (Sovik, 1993).

However, between 14 and 16 years (ninth grade of basic education to second year of high school- sophomore), speed tends to decrease, because the concern goes beyond calligraphy, because the requirements are related to planning and written text revision (Sampaio

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 349 of 404

and Cardoso, 2015; Kim et al., 2015), which are the most critical for writing performance (Bourdin and Fayol, 2002), therefore, ensuring a better written text - cohesive, coherent and respecting the standard language and communication (Alves et al., 2008; Pontart et al., 2013).

Conclusion

This chapter describes what is expected, in terms of writing proficiency, for each age group in the thematic Free Writing Task of DASH, being similar to a test at environment / examinations, for students, requiring legible writing, with sequential ideas within a certain period of time. The main results showed that:

- Age is not a determining factor for the calligraphic quality, but it influences the writing speed;
- The students tend to write faster during the first two minutes of the task;
- Rate decreases as task exposure time increases
- From the 6th minute of the task, the performance of the students is similar, that is, from that time, writing speed does not differ when groups are compared.
- From 9 to 13 years, writing speed increases, however, from 14 to 16 years the rate declines.

From these findings, it can be emphasized here the importance that the education professional exerts, not only for handwriting development, but also on the knowledge expected in terms of legibility and writing speed, because, from this, the professional will be able to monitor and identify the difficulties presented by the students and observe underperformance , in accordance to chronological age.

Among the limitations of this study, it is necessary to carry out new studies that aim to compare the performance of students in relation to sexual gender and, also, the validation of the instrument, as these investigations will provide greater credibility and confidence for the findings presented here.

Acknowledgment

We thank the São Paulo Research Foundation - FAPESP, and Coordination for the Improvement of Higher Education Personnel (CAPES) for financial support.

References

Accardo, A. P., Genna, M., and Borean, M. (2013). Development, maturation and learning influence on handwriting kinematics. *Human movement science*, 32(1):136–146.

Almeida, I. C. (2013). Avaliao do processo de escrita manual em crianas com pobre qualidade da caligrafia e boa qualidade da caligrafia. *Master Diss., Escola superior da sade do Alcoito, Santa Casa da Misericrdia de Lisboa.*

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 350 of 404

Alves, R. A., Castro, S. L., and Olive, T. (2008). Execution and pauses in writing narratives: Processing time, cognitive effort and typing skill. *International journal of psychology*, 43(6):969–979.

Barnett, A., Henderson, S., Scheib, B., and Schulz, J. (2007). *The Detailed Assessment of Speed of Handwriting (DASH). Manual*. Pearson Education.

Benbow, M. (1995). Principles and practices of teaching handwriting. *Hand function in the child*, pages 255–281.

Berninger, V. W., Abbott, R. D., Augsburger, A., and Garcia, N. (2009). Comparison of pen and keyboard transcription modes in children with and without learning disabilities. *Learning Disability Quarterly*, 32(3):123–141.

Bourdin, B. and Fayol, M. (2002). Even in adults, written production is still more costly than oral production. *International Journal of Psychology*, 37(4):219–227.

Cardoso, M. H., Henderson, S., and Capellini, S. A. (2014). Translation and cultural adaptation of brazilian detailed assessment of speed of handwriting: conceptual and semantic equivalence. *Audiology-Communication Research*, 19(4):321–326.

Chartrel, E. and Vinter, A. (2006). Rôle des informations visuelles dans la production de lettres cursives chez lenfant et ladulte. *L'Année psychologique*, 106(1):43–63.

de Almeida, P. H. T. Q., Sorensen, C. B. S., Magna, L. A., Cruz, D. M. C., and Ferrigno, I. S. V. (2013). Avaliação da escrita através da fotogrametria–estudo da preensão trípode dinâmica. *Revista de Terapia Ocupacional da Universidade de São Paulo*, 24(1):38–47.

Feder, K. P. and Majnemer, A. (2007). Handwriting development, competency, and intervention. *Developmental Medicine & Child Neurology*, 49(4):312–317.

Graham, S., Berninger, V. W., Abbott, R. D., Abbott, S. P., and Whitaker, D. (1997). Role of mechanics in composing of elementary school students: A new methodological approach. *Journal of educational psychology*, 89(1):170.

Graham, S., Struck, M., Santoro, J., and Berninger, V. W. (2006). Dimensions of good and poor handwriting legibility in first and second graders: Motor programs, visual–spatial arrangement, and letter formation parameter setting. *Developmental neuropsychology*, 29(1):43–60.

Kim, Y.-S., Al Otaiba, S., and Wanzek, J. (2015). Kindergarten predictors of third grade writing. *Learning and individual differences*, 37:27–37.

McCarney, D., Peters, L., Jackson, S., Thomas, M., and Kirby, A. (2013). Does poor handwriting conceal literacy potential in primary school children? *International Journal of Disability, Development and Education*, 60(2):105–118.

Medwell, J. and Wray, D. (2008). Handwriting–a forgotten language skill? *Language and Education*, 22(1):34–47.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 351 of 404

Olive, T. (2010). Methods, techniques, and tools for the on-line study of the writing process. *Writing: processes, tools and techniques*, pages 1–18.

O'Mahony, P., Dempsey, M., and Killeen, H. (2008). Handwriting speed: duration of testing period and relation to socio-economic disadvantage and handedness. *Occupational Therapy International*, 15(3):165–177.

Overvelde, A. and Hulstijn, W. (2011). Handwriting development in grade 2 and grade 3 primary school children with normal, at risk, or dysgraphic characteristics. *Research in developmental disabilities*, 32(2):540–548.

Paz-Villagrán, V., Danna, J., and Velay, J.-L. (2014). Lifts and stops in proficient and dysgraphic handwriting. *Human movement science*, 33:381–394.

Peverly, S. T. (2006). The importance of handwriting speed in adult writing. *Developmental Neuropsychology*, 29(1):197–216.

Piek, J. P., Dawson, L., Smith, L. M., and Gasson, N. (2008). The role of early fine and gross motor development on later motor and cognitive ability. *Human movement science*, 27(5):668–681.

Pontart, V., Bidet-Ildei, C., Lambert, E., Morisset, P., Flouret, L., and Alamargot, D. (2013). Influence of handwriting skills during spelling in primary and lower secondary grades. *Frontiers in psychology*, 4:818.

Prunty, M. M., Barnett, A. L., Wilmut, K., and Plumb, M. S. (2013). Handwriting speed in children with developmental coordination disorder: Are they really slower? *Research in developmental disabilities*, 34(9):2927–2936.

Rosenblum, S., Goldstand, S., and Parush, S. (2006). Relationships among biomechanical ergonomic factors, handwriting product quality, handwriting efficiency, and computerized handwriting process measures in children with and without handwriting difficulties. *American Journal of Occupational Therapy*, 60(1):28–39.

Rosenblum, S., Parush, S., and Weiss, P. L. (2003a). Computerized temporal handwriting characteristics of proficient and non-proficient handwriters. *American Journal of Occupational Therapy*, 57(2):129–138.

Rosenblum, S., Weiss, P. L., and Parush, S. (2003b). Product and process evaluation of handwriting difficulties. *Educational Psychology Review*, 15(1):41–81.

Sampaio, M. N. ; Cardoso, M. H. (2015). A produo textual e a interferłncia da legibilidade e velocidade. *In: Olga Valria Campana dos Anjos Andrade; Paola Matiko Martins Okuda; Simone Aparecida Capellini. (Org.). Tpicos em Transtornos de Aprendizagem - Parte IV. 1ed.Marlia: Fundepe*, IV(1):73–88.

Sassoon, R. (2004). *The art and science of handwriting*. Intellect Books.

Sovik, N. (1993). Development of children's writing performance: Some educational implications. *Motor development in early and later childhood: Longitudinal approaches*, pages 229–246.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 352 of 404

Sumner, E., Connelly, V., and Barnett, A. L. (2013). Children with dyslexia are slow writers because they pause more often and not because they are slow at handwriting execution. *Reading and Writing*, 26(6):991–1008.

Tomchek, S. and Schneck, C. (2006). Evaluation of handwriting. *Hand function in the child: Foundations for remediation*, pages 291–309.

Tucha, O., Tucha, L., and Lange, K. W. (2008). Graphonomics, automaticity and handwriting assessment. *Literacy*, 42(3):145–155.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 353 of 404

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

In: Handwriting: Recognition, Development and Analysis ISBN: 978-1-53611-937-4
Editors: Byron L. D. Bezerra et al. © 2017 Nova Science Publishers, Inc.

Chapter 14

DATASETS FOR HANDWRITTEN SIGNATURE VERIFICATION: A SURVEY AND A NEW DATASET, THE RPPDI-SIGDATA

***Victor Kléber Santos Leite Melo*[1,*], *Byron Leite Dantas Bezerra*[1,†], *Rebecca H. S. N. Do Nascimento*[1,‡], *Gabriel Calazans Duarte de Moura*[1,§], *Giovanni L. L. de S. Martins*[1,¶], *Giuseppe Pirlo*[2,‖] *and Donato Impedovo*[2,**]**

[1]Polytechnic School, University of Pernambuco, Brazil
[2] Dipartimento di Informatica,
Università degli Studi di Bari, Italy

1. Introduction

In the modern society, biometrics technology is used in several security applications for personal authentication. The aim of such systems is to confirm the person identity based on physiological or behavioral traits. In the first case, recognition is based on biological characteristics such as fingerprint, palmprint, iris, face, etc. The latter relies on behavioral traits such as voice pattern and handwritten signature (Jain et al., 2004).

Handwritten signature remains as one of the main approaches for identity authentication. One of the reasons for its widespread is because the acquisition is easy, non-invasive and most individuals are familiar with its use in their daily life (Impedovo and Pirlo, 2008). Due to its convenient nature, signatures can be employed as a sign of confirmation in a wide set of important documents, especially on bank checks, credit card transactions, identification documents and a variety of business certificates and contracts.

*E-mail address: vkslm@ecomp.poli.br.
†E-mail address: byronleite@ecomp.poli.br.
‡E-mail address: rhsnn@ecomp.poli.br.
§E-mail address: gcdm@ecomp.poli.br.
¶E-mail address: gllsm@ecomp.poli.br.
‖E-mail address: giuseppe.pirlo@uniba.it.
**E-mail address: donato.impedovo@uniba.it.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

However, as a behavioral trait, signatures are susceptible to spoof attacks, which is the attempt to spoof the signature of an enrolled user to fool the system (Jain et al., 2004). Two types of impostors are considered, specifically: casual impostors (producing random forgeries) when no information about authentic writer signature is known, and real impostors (producing skilled forgeries) when some information of the signature is used (Fierrez and Ortega-Garcia, 2008).

If a signature on a document is false, this document is also considered invalid. Thus, prevent frauds in the signature verification process has been a challenge for researchers around the world. However, manual signature-based authentication of a large set of documents is difficult and also a time-consuming and labor-intensive task. Hence, several Automatic Handwritten Signature Verification Systems (AHSVS) have been proposed. These systems aim to automatically decide if a query signature is in fact of a particular person or not.

AHSVS are essentially a pattern recognition application that works by receiving a signature as input, extracting a feature set from the data and classifying the sample using a template database as a reference. As any pattern recognition system, AHSVS are learning-based, which requires a dataset that can be used to assess their performances to create accurate signature verification methods. These datasets contain signatures digitized by either using an optical scanner to obtain the signature directly from the paper or by using an acquisition device such as digitizing tablets or electronic pens with digital ink. The two approaches are identified as offline (static) and online (dynamic), respectively. In the online modality, data is stored during the writing process and consists of a temporal sequence of the two-dimensional coordinates (x, y) of consecutive points, whereas in the offline case, only a static representation of the completed writing is available as an image. Moreover, each representation has specific attributes not present in the other (Viard-Gaudin et al., 1999). For instance, online data do not include information about the width of the strokes and the texture of the ink on the paper, while the offline representation has lost all dynamic information of the writing process. As a result, features such as pen trajectory, which can be easily computed in the online domain, can only be inferred from a static image (Nel et al., 2005).

In the last few years, several handwritten signature datasets have been created and some made publicly available. The general corpus consists of a set of genuine and forgery signatures for each writer and can be categorized on different dimensions including the modality (online or offline), script and size.

Although many datasets containing samples for offline, online or both modalities combined haven been proposed, those datasets normally do not convey some important real world challenges, not assessing the robustness of the systems on real-world scenarios. Consequently, said systems often fail to deliver the expected results when employed in practice (Ahmed et al., 2013).

In practical scenarios, signatures are acquired on a wide set of conditions and in both modalities. Different conditions for online acquisition includes signatures acquired in several types of devices, e.g., using smartphones or different models of digitizing tablets. Moreover, when dealing with offline signatures, most of the samples are present in documents with complex backgrounds and with different signing area constraints. Examples of such documents include bank checks, contracts, identification documents, forms, etc.

Cerence Operating Company; Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 356 of 404

(Ahmed et al., 2013, 2012). Those distinct types of signatures often need to be integrated into the same system in an interoperable manner.

In regards to signature verification interoperability, many research problems are open to investigation, such as (i) development of complete document authentication systems involving both signature segmentation and verification process taking into account different signing area constraints (ii) analysis of the implications on AHSVS of the combination of signatures acquired on smartphones and conventional digitizing tablets (iii) development of systems capable of integrating both online and offline samples interchangeably, towards a unified signature biometry. With the currently available datasets, investigation on the direction of the listed research problems is limited to samples acquired in controlled environments or can not be made at all.

Works have been done on topics directed towards signature verification interoperability. Qiao et al. (2007) proposed an offline signature verification system that uses online handwriting signatures instead of images in the registration phase, however, in the experiments the authors used synthetic offline images generated by the interpolation of online signature samples. Uppalapati (2007) proposed a system to integrate both modalities of handwritten signatures, not only providing a method to match offline signatures against an online and vice-versa, but also using both static and dynamic features, when available, to improve the system performance. Ahmed et al. (2012) proposed a method for signature extraction from documents, it is noteworthy that the segmentation accuracy was evaluated only on the patch level. According to the authors, it is due to the lack of publicly available datasets containing the ground truth of signatures on the stroke level. Ahmed et al. (2013) discuss the currently non-applicability of most signature verification systems and the lack of complete document authentication systems involving signature segmentation and verification. According to the authors, it is due to the absence of datasets suitable for the development of such systems, containing both patch and stroke level ground truth. Pirlo et al. (2015) investigated the effects of signing area constraints on geometric features of online signatures. Diaz-Cabrera et al. (2014) proposed several approaches to synthetically generate offline signatures simulating the pen ink deposition on the paper based on dynamic information from online signatures. Zareen and Jabin (2016) presented a comprehensive survey of mobile-biometric systems and proposed a method for online signature verification. The approach was evaluated on the SVC (Yeung et al., 2004) dataset and a database acquired using a mobile device (Martinez-Diaz et al., 2014).

Aiming to overcome the limitations of the current state of handwritten signature datasets, we present the RPPDI-SigData, an evaluation dataset for AHSVS that includes signatures captured for both online and offline modalities and from different signing conditions. Samples for the online modality were acquired on smartphones and digitizing tablets and for the offline domain acquired in documents with complex backgrounds (including a stroke level ground-truth) and different signing area constraints.

Alongside with the description of the RPPDI-SigData, this chapter also summarizes 17 publicly available handwritten signature datasets. Our goal is to provide the reader an overview of the existing evaluation datasets and its main characteristics such as the number of samples, protocols, type of forgeries, script.

The rest of the chapter is structured as follows: Section 2 presents an overview of the existing evaluation datasets for handwritten signature verification. Section 3 describes

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 357 of 404

RPPDI-SigData, our proposed dataset and we conclude the chapter in Section 4.

2. Handwritten Signature Datasets

In this section, we present 17 publicly available datasets used in AHSVS. We have focused our selection on datasets that are public and also have a documented protocol. The datasets included in this review are used for offline/online signature verification and also signature segmentation from documents.

2.1. GPDS960

The GPDS-960 signature corpus (Vargas et al., 2007) is a Spanish offline handwritten signature database containing genuine and forgery samples for 960 individuals. For each subject there are 24 genuine signatures and 30 forgeries, producing 960 x 24 = 23049 genuine signatures and 960 x 30 = 28800 forgeries. The genuine signatures were taken in one session which the subjects were asked to sign his/her signature on a form with 24 signing boxes, which half the boxes were 5x3.5cm and the other half 5.5x2.5cm. The forgeries were made by a different set of 1920 individuals, each forger filled up a form containing 15 boxes, specifically for 5 randomly selected genuine signatures that should be forged 3 times. Each form was scanned with an HP4400 device using 256 level grayscale and 300 dpi resolution. There is a black and white and gray scale version of the dataset, however, due to a problem during a move, the authors lost the gray scale signatures of 79 users, remaining 881 sets of writers in the gray scale version of this dataset.

2.2. MCYT

The MCYT bimodal biometric database consists of a collection of fingerprint and signature samples for 330 contributors from 4 different Spanish sites. Specifically for the MCYT-Signature subcorpus, the samples were acquired using a pen tablet, model INTUOS A6 USB. During the acquisition, users provided their signatures using a pen and a paper template over the tablet enabling both online as the offline sample to be captured. The creators made available two subsets of the MCYT-Signature, namely the MCYT-100 (Ortega-Garcia et al., 2003) and MCYT-75 (Fierrez-Aguilar et al., 2004). The first contains 25 genuine and 25 forgeries online samples of 100 authors, and the latter 15 genuine and 15 forgeries offline signatures of 75 writers.

2.3. UTSig

The UTSig (University of Tehran Persian Signature) dataset (Soleimani et al., 2016) is a Persian offline handwritten signature dataset consisting of samples for 115 male subjects. Each subject has 27 genuine and 45 forged specimens of his signature, we can see in Figure 1 samples of 4 writers. Genuine signatures were acquired using a form containing 10 signing areas of 6 different sizes. The acquisition was made in 3 days, on each day the writers signed 9 genuine signatures and the last the subject signed with his opposite hand. They collected 3105 genuine and 345 opposite-hand signatures. The forged samples of the dataset are divided into 3 categories.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 358 of 404

The creators consider as the first category of forgeries the 345 opposite-hand signatures made by the authentic writers. The second category contains forged samples obtained from a different set of 230 subjects. They were asked to fill 3 forms each containing 6 signing boxes for 3 different writers, the forgers were free to practice as much as they want and the observable sample varied from one to three genuine signatures of the authentic writer. The last category was made by a more skilled forger, the form was the same as the second category and the observable sample was only one genuine signature. All forms were scanned with 600 dpi resolution and stored as 8-bit grayscale TIF files.



Figure 1. Four genuine samples and their respective forgeries of the UTSig dataset.

The BIOMET (Garcia-Salicetti et al., 2003) is a multimodal database including data from five different modalities: fingerprint, face, voice, hand and online handwritten signatures. The dataset was collected in 3 different sessions with 130, 106 and 91 participants, hence a total of 327 contributors. For each subject, there are 15 genuine and 17 forgery signatures. The data stored for each sample were x-y coordinates, the pressure, azimuth and altitude of the pen. The device used during the acquisition was a Wacom Intuos2 tablet with different pens. In the first session a Grip Pen was used, in the second and third session an Ink Pen was used, leading to a more natural writing, since the subject signs in a sheet of paper over the tablet. The forgeries were made by 5 individuals different from the authentic writers.

2.4. MyIDea

MyIDea (Dumas et al., 2005) is a multimodal biometric dataset that includes traits such as talking face, audio, fingerprints, signature, handwriting and hand geometry. In particular, the handwritten signatures were acquired using an A4 Intuos2 graphic tablet from Wacom, the tablet records 5 parameters: x-y coordinates, pressure, azimuth, and altitude. An Intuos

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 359 of 404

InkPen is used to write on a paper over the tablet, the main advantage of this procedure is to record both online and offline data at the same time, ensuring a natural writing using an almost standard pen and paper. The dataset contains data from 73 users from whom 46 users are French and 27 English. Each subject is represented with 18 genuine signatures and 36 skilled forgeries. The forgeries were captured in two conditions. In the first condition, the forgers were asked to produce 18 samples using only a paper-copy of the authentic signature as a reference. For the second condition, the forgers were able to use a dedicated software to study the signature dynamics on screen. In both conditions, they were allowed to train as long as they want to gain confidence on the authentic signature format.

2.5. SIGMA

SIGMA (Ahmad et al., 2008) is a dataset containing signatures for 213 individuals. For each subject, 30 genuine signatures were collected, and 10 forgeries were produced. During the acquisition of the forgeries, the subjects were given as much time and number of genuine samples as they want. The creators also checked the resemblance of the forgeries to the original signatures before including the sample in the database. The acquisition was made using an A4 Intuos3 graphic tablet from Wacom, the online data stored was the x-y coordinates along with the pen pressure at each coordinate. The offline sample was also captured using an electronic ink pen on a paper form shown on Figure 2. Offline forms were scanned as grayscale images using a 600 dpi resolution scanner.

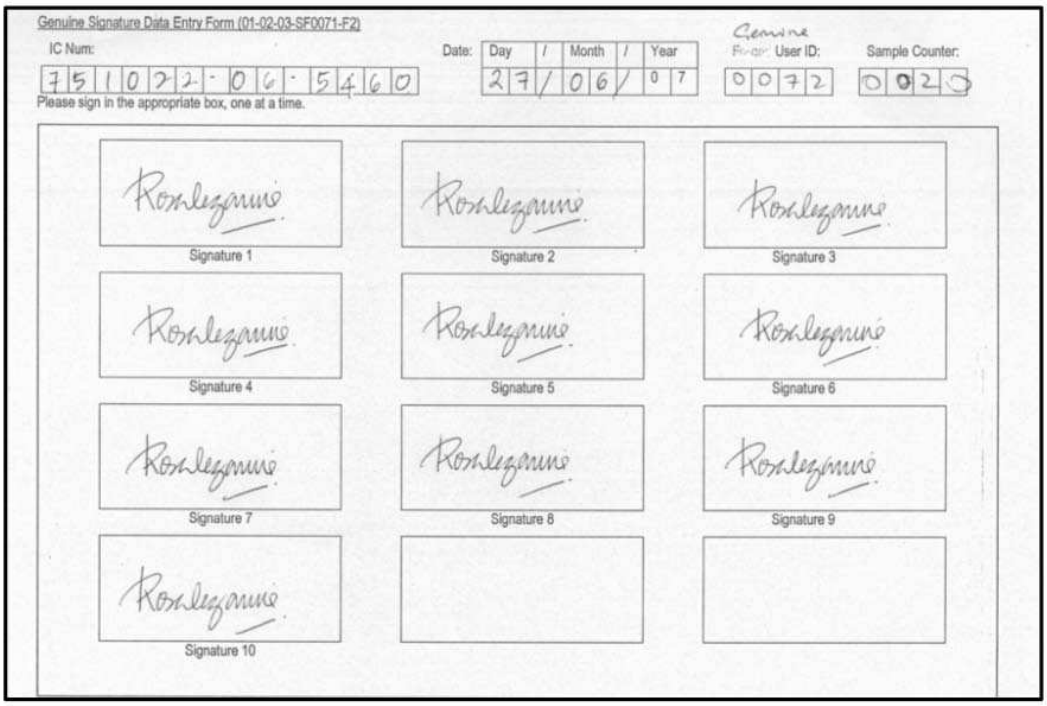


Figure 2. Form used on the signature acquisition of the SIGMA dataset.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

2.6. SigComp2009

ICDAR 2009 Signature Verification Competition (SigComp2009) (van den Heuvel et al., 2009) is a signature verification competition, which provided an online and offline dataset containing training and evaluation sets. For training, NISDCC signature collection acquired in WANDA project (Franke et al., 2003) was used. The training dataset consists of 12 users, for each subject 5 genuine signatures were collected and 5 forged samples were made. The forgeries samples were made by 31 forgers. The test set was collected in the Netherlands Forensic Institute (NFI). It contains data for 100 writers and each represented by 12 genuine and 6 forged samples made by the other participants.

2.7. 4NSigComp2010

The 4NSigComp2010 comprises two scenarios, the dataset for the first scenario was collected by La Trobe University and was made available with the competition. whereas the second scenario is a subset of the GPDS-960 database, previously reported on this chapter. Signatures for the first scenario were collected using a ball-point pen on paper and scanned at 600 dpi.

The training set of the first scenario consists of 9 reference signatures by one author and 200 questioned signatures including 76 genuine, 104 simulated (forged) signatures made by 27 freehand forgers, and 20 disguised samples. Genuine and disguised samples were made over a week. In addition, the writer signed another 81 genuine signatures to be used as a reference for the forgery acquisition. The forgeries were made with the contribution of 27 volunteers, each subject had 3 out of 81 genuine samples, and imitated without tracing in two ways: inspect the genuine signature and forge 3 times without practice and practice simulating the genuine signature 15 times then forge the signature 3 times. The test set has 25 signatures of another person written during 5 days and 100 questioned samples including 3 genuine, 7 disguise, and 90 simulated (forged) signatures written by 34 freehand forgers whom were either laypersons or calligraphers.

2.8. SigComp11

SigComp2011 (Liwicki et al., 2011) includes a dataset containing Chinese and Dutch signatures samples. The collection contains both offline and online signature samples. The acquisition of the signatures was made using a Wacom Intuos3 A3 Wide USB Pen Tablet, the paper placed over the tablet had 12 signing boxes of size 59 x 23 mm. The paper was scanned at a 400 dpi, RGB color and stored as eps images. Online data stored includes x-z coordinates and pressure. Due to some problems during the signature acquisition, a number of signatures in the online data sets are different from those in the offline set, an overview of the number of samples on both datasets and both modes are provided in Table 1

2.9. 4NSigComp2012

In 4NSigComp2012 (Liwicki et al., 2012) the dataset introduced is similar to the database of Scenario 1 of 4NSigComp2010. The data collected contains 3 sets of authentic authors, A1, A2, and A3 respectively. Table 2 shows the number of specimens for each category of signature samples.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 361 of 404

Table 1. Overview of the amount of samples provided in the SigComp11 dataset. Number of authors (A) and number of genuine (G) (Reference (GR) and questioned (GQ)) and forged (F).

Dataset	Training			Test			
	A	G	F	A	GR	GQ	F
Chinese Offline	10	235	340	10	116	120	367
Chinese Online	10	230	430	10	120	125	461
Dutch Offline	10	240	123	54	648	648	638
Dutch Online	10	330	119	54	648	648	611

Table 2. Overview of the dataset introduced in the 4NSigComp2012.

Type of signature		Author A1	Author A2	Author A3
Reference		20	16	15
Questioned	**Disguised**	47	08	09
	Forged	160	42	71
	Genuine	43	50	20

Genuine and disguised samples were collected during 10 to 15 days. The number of forgers varied from 2 to 31. Each forger was provided with 3 to 6 authentic samples. Forgers used pen and pencil and forged with and without practice.

2.10. SigWiComp13

In the SigWiComp13 (Malik et al., 2013) an offline Dutch and an online and offline Japanese signature datasets were introduced. The offline Japanese dataset was converted from online signatures that contain 30 classes with 42 genuine samples per class, made in 4 days, and 36 forgeries made by 4 forgers. Dutch dataset has 27 authentic writers who made 10 signatures with arbitrary writing instruments during 5 days. For forgeries, 9 subjects used any to all of the supplied specimen signatures as a reference. In average, there are 36 forgeries for each authentic writer.

2.11. SigWiComp2015

With the SigWiComp2015 (Malik et al., 2015) three datasets were made available, namely Bengali, Italian and German datasets. The Italian dataset consists of offline binarized signatures divided into two separated sets: training and testing set. The training set is composed of 50 writers with 5 reference signatures each. The testing set contains samples for the same 50 authors with 10 questioned signatures, each being either a genuine or a skilled forgery signature. During the acquisition of the forgeries, the subjects were allowed to practice as much as they want.

The Bengali offline signatures were collected from 10 contributors, each providing 24 genuine signatures. The imitators made 30 forgeries for each authentic writer and during the process they were allowed to practice. All images were captured at 300 dpi resolution with 256 levels of gray.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 362 of 404

The German online signatures were collected using the Anoto digitizing pen (Malik et al., 2012) rather than a tablet. Data stored in this dataset includes x-y coordinates and pressure. For the training set, 30 genuine authors provided 10 genuine signatures, for the evaluation, data for the same 30 authors containing 15 questioned specimens including genuine and skilled forgeries samples.

2.12. SUSIG

The SUSIG (Sabanci University Signature) (Kholmatov and Yanikoglu, 2009) is divided into two subcorpora: visual and blind. Signatures in the Visual Subcorpus were acquired using an Interlink Electronics's ePad-ink tablet which has a pressure-sensitive LCD screen such that subjects could see their signatures while signing, whereas the Blind subcorpus was collected using Wacom's Graphire2 tablet and pen which provides no visual feedback. Data stored for both subcorpora contains the x-y coordinates, time stamp, and pressure level for each point.

The Visual Subcorpus consists of signatures of 100 individuals acquired over two sessions that were approximately one week apart. Each subject provided 10 samples of his/her signature on each session, for a total of 20 genuine samples. Each individual was asked to forge a randomly selected signature. Two types of forgeries were considered: skilled and very skilled, 5 for each type. Regarding the first type, the forger could watch the signature's animation on a monitor, whereas in the second type the animation of the genuine reference was also mapped on the LCD screen of the tablet in such a way that the forger could trace over.

The Blind Subcorpus contains signatures of a different set of 100 individuals, a group of 30 subjects provided 8 genuine signatures while the rest of the 70 supplied 10 genuine signatures each. Each user provided 10 forgeries for a randomly selected writer and the forgery acquisition process was the same as the skilled forgery of the Visual Subcorpus.

2.13. SVC2004

SVC2004 (Yeung et al., 2004) was the first international signature verification competition. The competition was divided into two separate tasks using two different signature databases of online handwritten signatures. The difference between the two tasks is the information collected for the dataset. While in the first task the data collected contains only coordinate information, data for the second task also contains additional information such as pressure and pen orientation. The aim of different tasks is to simulate acquisition on devices such as personal digital assistants (PDA) in the case of the first task and digitizing tablets on the second task. Each dataset contains a set of signatures for 100 subjects, each set is composed of 20 genuine signatures and 20 skilled forgeries made by at least four other contributors.

2.14. CEDAR

The CEDAR (Kalera et al., 2004) dataset is an offline handwritten signature dataset with 55 signers and a total of 2640 signature samples. For each signer, 24 genuine signatures were collected and 20 arbitrary signers were chosen to skillfully forge the genuine specimens, producing 24 skilled forgeries for each subject of the database. The signing area of the

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 363 of 404

form used to collect the signatures is a predefined space of 2 x 2 inches. The forms were scanned at 300 dpi in 256 levels of gray and stored in eps format.

2.15. SID-Signature Database

SID-Signature Database (Abdelghani and Amara, 2013) is a Tunisian offline signature dataset. The corpus consists of a set of 100 subjects, each containing 60 genuine, 20 simple forgeries and 20 skilled-imitated forgery sample. The form used to acquire the signatures was an A4 paper divided into 15 signing areas of the same size. The genuine samples were acquired in different sessions to avoid variations related to fatigue. A group of 10 people different from the authentic writers were asked to create 2 simple and 2 skilled imitation forgeries of each authentic writer. The creators of the dataset consider a simple forge writing the surname and/or the name of the genuine writer, whereas during the skilled imitation they provided the forger different static images of the genuine signature, however, the forgers could not practice the imitation. The forms have been digitized at a 300 dpi resolution with 256 gray levels and they used three scanners: an HP 3200C, HP G2710 and EPSON DX4400. The creators preprocessed the forms and stored the signature images as a black and white bitmap file.

2.16. Tobacco800

Tobacco800 (tob, 2007) is a dataset composed of 1290 complex document images which contain information about signatures on printed text documents. Resolutions of documents in Tobacco800 vary significantly from 150 to 300 DPI, and the dimensions of images range from 1200 by 1600 to 2500 by 3200 pixels. The dataset also includes a patch level ground truth for the signatures. Figure 3 shows a document containing a signature extracted from the dataset.

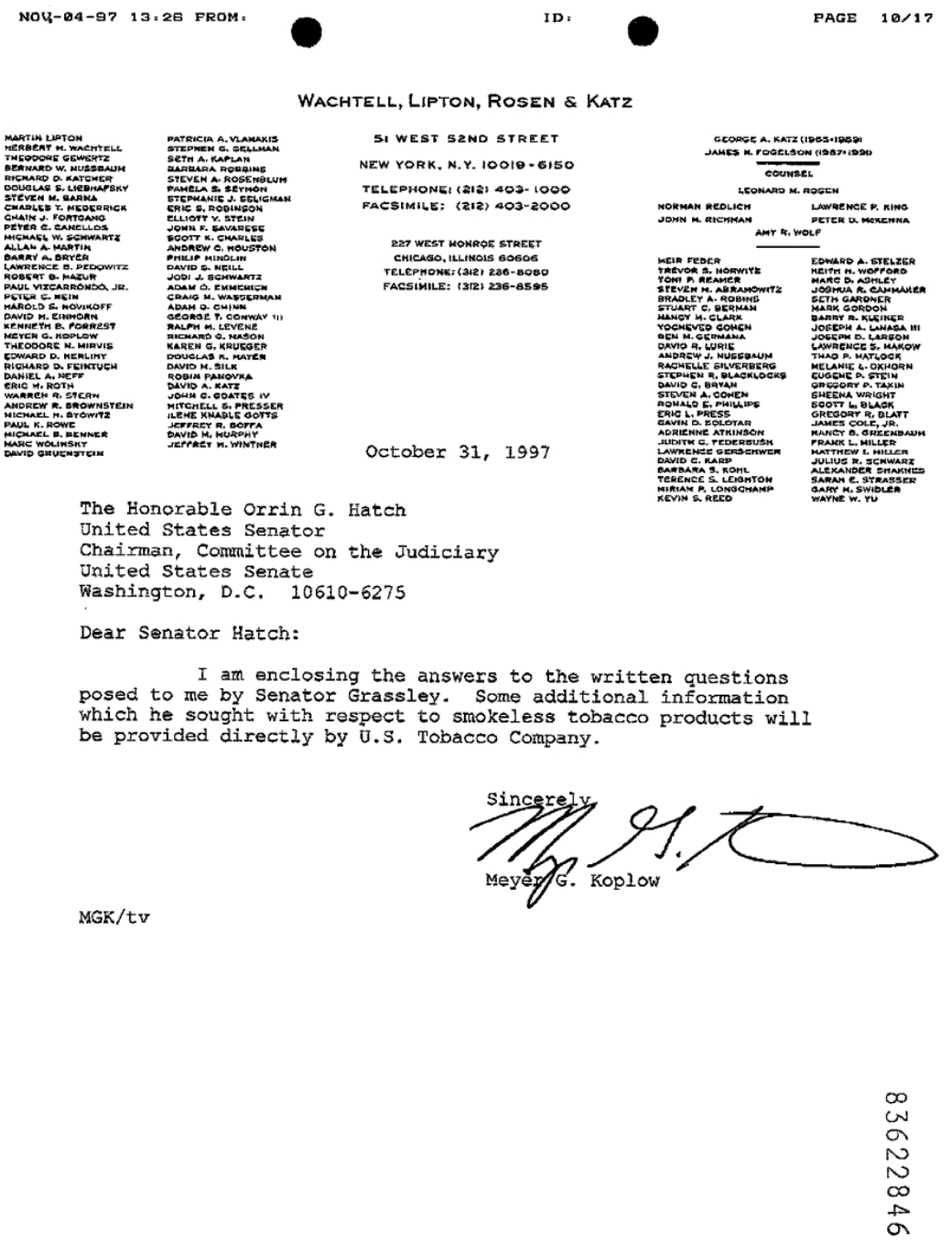
3. RPPDI-SigData Dataset

In this section, we describe our proposed dataset, RPPDI-SigData. The main contribution of this new dataset is the effort to complement existing handwritten signature verification datasets towards closing the gap between the AHSVS research and real-world applications. This evaluation dataset provides signature samples acquired on both signature modalities (online and offline).

During the acquisition of the online samples, two devices were used, namely, a Wacom STU-530 digitizing tablet and a Samsung Galaxy Note, in both cases using the specific stylus for each device. Signatures of the offline modality were collected on 4 different documents and a form containing 4 different signing area sizes. Alongside the signed documents it is also included a stroke level ground truth.

The purpose of providing signature samples using different signature acquisition devices and with different document formats is to support the performance assessment for signature verification interoperability. For instance, researchers can, but are not restricted to, use this dataset to evaluate (i) complete document authentication systems that employ

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 364 of 404



NOV-04-97 13:26 FROM: ID: PAGE 10/17

WACHTELL, LIPTON, ROSEN & KATZ

MARTIN LIPTON
HERBERT M. WACHTELL
THEODORE GEWERTZ
BERNARD W. NUSSBAUM
RICHARD D. KATCHER
DOUGLAS S. LIEBHAFSKY
STEVEN M. GARNA
CHARLES T. MEDERRICK
CHAIN J. FORTGANG
PETER C. CANELLOS
MICHAEL W. SCHWARTZ
ALLAN A. MARTIN
BARRY A. BRYER
LAWRENCE B. PEDOWITZ
ROBERT B. MAZUR
PAUL VIZCARRONDO, JR.
PETER C. HEIN
HAROLD S. NOVIKOFF
DAVID M. EINHORN
KENNETH B. FORREST
MEYER G. KOPLOW
THEODORE N. MIRVIS
EDWARD D. HERLIHY
RICHARD D. FEINTUCH
DANIEL A. NEFF
ERIC M. ROTH
WARREN R. STERN
ANDREW R. BROWNSTEIN
MICHAEL H. BYOWITZ
PAUL K. ROWE
MICHAEL B. RENNER
MARC WOLINSKY
DAVID GRUENSTEIN

PATRICIA A. VLAHAKIS
STEPHEN G. SELLMAN
SETH A. KAPLAN
BARBARA ROBBINS
STEVEN A. ROSENBLUM
PAMELA S. SEYMON
STEPHANIE J. SELIGMAN
ERIC S. ROBINSON
ELLIOTT V. STEIN
JOHN F. SAVARESE
SCOTT K. CHARLES
ANDREW C. HOUSTON
PHILIP MINDLIN
DAVID G. NEILL
JODI J. SCHWARTZ
ADAM O. EMMERICH
CRAIG M. WASSERMAN
ADAM O. CHINN
GEORGE T. CONWAY III
RALPH M. LEVENE
RICHARD G. MASON
KAREN G. KRUEGER
DOUGLAS K. MAYER
DAVID H. SILK
ROBIN PANOVKA
DAVID A. KATZ
JOHN C. COATES IV
MITCHELL S. PRESSER
ILENE KNABLE GOTTS
JEFFREY R. BOFFA
DAVID M. MURPHY
JEFFREY M. WINTNER

51 WEST 52ND STREET
NEW YORK, N.Y. 10019-6150
TELEPHONE: (212) 403-1000
FACSIMILE: (212) 403-2000

227 WEST MONROE STREET
CHICAGO, ILLINOIS 60606
TELEPHONE: (312) 236-8080
FACSIMILE: (312) 236-8595

GEORGE A. KATZ (1955-1969)
JAMES H. FOGELSON (1957-1990)

COUNSEL
LEONARD M. ROSEN
NORMAN REDLICH
JOHN M. RICHMAN
LAWRENCE P. KING
PETER D. MCKENNA
AMY R. WOLF

MEIR FEDER
TREVOR S. NORWITZ
YONI P. REAMER
STEVEN M. ABRAMOWITZ
BRADLEY A. ROBINS
STUART C. BERMAN
NANCY H. CLARK
YOCHEVED COHEN
BEN M. GERMANA
DAVID A. LURIE
ANDREW J. NUSSBAUM
RACHELLE SILVERBERG
STEPHEN R. BLACKLOCKS
DAVID C. BRYAN
STEVEN A. COHEN
RONALD E. PHILLIPS
ERIC L. PRESS
GAVIN D. SOLOTAR
ADRIENNE ATKINSON
JUDITH G. FEDERBUSH
LAWRENCE GERSCHWER
DAVID C. KARP
BARBARA S. KOHL
TERENCE S. LEIGHTON
MIRIAN P. LONGCHAMP
KEVIN S. REED

EDWARD A. STELZER
KEITH N. WOFFORD
MARC D. ASHLEY
JOSHUA R. CAMMAKER
SETH GARDNER
MARK GORDON
BARRY R. KLEINER
JOSEPH A. LANASA III
JOSEPH D. LARSON
LAWRENCE S. MAKOW
THAD P. MATLOCK
MELANIE L. OKHORN
EUGENE P. STEIN
GREGORY P. TAXIN
SHEENA WRIGHT
SCOTT W. BLACK
GREGORY R. BLATT
JAMES COLE, JR.
NANCY B. GREENBAUM
FRANK L. MILLER
MATTHEW I. MILLER
JULIUS M. SCHWARZ
ALEXANDER SHAKNES
SARAH E. STRASSER
GARY M. SWIDLER
WAYNE W. YU

October 31, 1997

The Honorable Orrin G. Hatch
United States Senator
Chairman, Committee on the Judiciary
United States Senate
Washington, D.C. 10610-6275

Dear Senator Hatch:

I am enclosing the answers to the written questions posed to me by Senator Grassley. Some additional information which he sought with respect to smokeless tobacco products will be provided directly by U.S. Tobacco Company.

Sincerely,

Meyer G. Koplow

MGK/tv

83622846

Figure 3. Document containing a signature extracted from the Tobacco800 dataset.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 365 of 404

both signature segmentation and verification methods (ii) the integration of signature verification in different acquisition devices using signatures acquired on a smartphone and a conventional digitizing tablet (iii) systems capable of integrating both online and offline samples interchangeable. Those open research problems are currently hardly addressed on the current state of publicly available datasets, please see Section 2

Regarding the composition of the dataset, it consists of signature samples for 15 contributors, for each contributor an Offline set and Online set were collected. The offline set includes 1 signature on the ID, voter ID, driver's license and check. Also, a filled signature frame of 8 signatures, hence 12 offline samples. The online set includes 6 smartphone signatures and 4 graphical tablet signatures, thus 10 online signatures. The forgeries were made in 2 checks, one filled signature frame (8 signatures), 6 on smartphones and 4 on a digitizing tablet, hence 10 offline and 10 online forgeries. Table 3 summarizes the number of samples collected in each category of the dataset.

Table 3. Overview of the quantity of samples collected in the RPPDI-SigData dataset. Information provided for number of signed ID documents (ID), driver's license (DL), voter ID (VID), checks (C) and signature frame (SF) and also the amount of online samples.

Type	Offline					Online	
	ID	DL	VID	C	SF	Smartphone	Pressure-tablet
Genuine	1	1	1	1	8	6	4
Forgery	0	0	0	2	8	6	4

The dataset was collected at the Polytechnic School of the University of Pernambuco, the majority of the contributors were students. Participants were briefly informed about the purpose of the acquisition but were not given details on how a signature verification system works.

The genuine offline signatures were collected using personal documents such as ID, voter ID and driver's license, a check and a preprinted form containing 8 signing boxes of 4 different sizes. All the documents were scanned at 300 dpi and stored as RGB files. Figure 4 shows an example of those documents filled by a contributor. The online samples were acquired using the Wacom STU-530 pressure tablet with a special type of pen and also on a smartphone Samsung Galaxy Note, using an appropriate stylus. The data stored in both devices was x-y coordinates, pressure and time.

The forgeries collected for the dataset were made by the participants. The forger could take as long as they want to practice the signature and use all samples of the authentic writer. For the evaluation of the process of extraction, the dataset contains a stroke level ground truth of the signature on the documents. We followed a semi-automatic procedure to create those images. Figure 5 shows a sample of ground truth in the dataset for a signed ID document.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 366 of 404



Figure 4. Example of each document used for the signature acquisition.

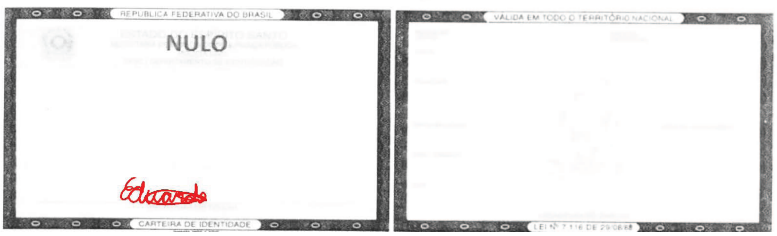


Figure 5. A sample of a groundtruth extracted from the dataset.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 367 of 404

Table 4. Summary of the public available datasets discussed on this review

Dataset / Ref	Year	Modality
GPDS960 (Vargas et al., 2007)	2007	Offline
BIOMET (Garcia-Salicetti et al., 2003)	2003	On/Off
SID - Signature Database (Abdelghani and Amara, 2013)	2013	Offline
MCYT-75 (Fierrez-Aguilar et al., 2004)	2004	Offline
4NSigComp2010 Scenario 1 (Liwicki et al., 2010)	2010	Offline
UTSig (Soleimani et al., 2016)	2016	Offline
CEDAR (Kalera et al., 2004)	2009	Offline
MyIDea (Dumas et al., 2005)	2005	On/Off
SigComp09 (van den Heuvel et al., 2009)	2009	On/Off
SIGMA (Ahmad et al., 2008)	2008	On/Off
SigComp11 (Liwicki et al., 2011)	2011	On/Off
SigWiComp13 (Malik et al., 2013)	2013	On/Off
SigWIComp2015 (Malik et al., 2015)	2015	On/Off
SUSIG (Kholmatov and Yanikoglu, 2009)	2009	Online
MCYT-100 (Ortega-Garcia et al., 2003)	2003	Online
SVC 2004 (Yeung et al., 2004)	2004	Online
Tobacco 800 (tob, 2007)	2007	Segmentation

Conclusion

In this chapter, we provide an overview of 17 publicly available handwritten signature datasets, which are summarized in Table 4. Based on our review, we found the lack of datasets containing signatures which were not pre-segmented from documents and enables multi-domain investigations. This motivated us to build a new dataset, RPPDI-SigData, which allows the integrated process of signature extraction and verification of samples acquired in different types of documents with complex backgrounds. The dataset also consists of online signatures acquired from a mobile device and a conventional digitizing tablet.

With the availability of our proposed dataset, we can address signature verification interoperability. We plan in future works using this dataset to evaluate a system that verifies signatures in different documents and acquisitions sources.

Acknowledgments

The authors would like to thank the CNPQ for supporting the development of this chapter through the research projects granted by “Edital Universal” (Process 444745/2014-9) and “Bolsa de Produtividade DT” (Process 311912/2015-0). In addition, the authors also acknowledge the Document Solutions for providing some of the devices used in this research.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 368 of 404

References

(2007). *The Legacy Tobacco Document Library (LTDL)*. University of California, San Francisco.

Abdelghani, I. A. B. and Amara, N. E. B. (2013). SID Signature Database: A Tunisian Off-line Handwritten Signature Database. In *International Conference on Image Analysis and Processing*, pages 131–139. Springer.

Ahmad, S. M. S., Shakil, A., Ahmad, A. R., Balbed, M. A. M., and Anwar, R. M. (2008). SIGMA-A Malaysian signatures database. In *2008 IEEE/ACS International Conference on Computer Systems and Applications*, pages 919–920. IEEE.

Ahmed, S., Malik, M. I., Liwicki, M., and Dengel, A. (2012). Signature segmentation from document images. In *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*, pages 425–429. IEEE.

Ahmed, S., Malik, M. I., Liwicki, M., and Dengel, A. (2013). Towards Signature Segmentation & Verification in Real World Applications. In *Proceedings of the 16th Biennial Conference of the International Graphonomics Society*, page 139–142.

Diaz-Cabrera, M., Gomez-Barrero, M., Morales, A., Ferrer, M. A., and Galbally, J. (2014). Generation of enhanced synthetic off-line signatures based on real on-line data. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 482–487. IEEE.

Dumas, B., Pugin, C., Hennebert, J., Petrovska-Delacrétaz, D., Humm, A., Evéquoz, F., Ingold, R., and Von Rotz, D. (2005). MyIDea-multimodal biometrics database, description of acquisition protocols. *Proc. Third COST*, 275:59–62.

Fierrez, J. and Ortega-Garcia, J. (2008). On-line signature verification. In *Handbook of biometrics*, pages 189–209. Springer.

Fierrez-Aguilar, J., Alonso-Hermira, N., Moreno-Marquez, G., and Ortega-Garcia, J. (2004). An off-line signature verification system based on fusion of local and global information. In *International Workshop on Biometric Authentication*, pages 295–306. Springer.

Franke, K., Schomaker, L., Veenhuis, C., Taubenheim, C., Guyon, I., Vuurpijl, L., van Erp, M., and Zwarts, G. (2003). WANDA: A generic Framework applied in Forensic Handwriting Analysis and Writer Identification. *HIS*, 105:927–938.

Garcia-Salicetti, S., Beumier, C., Chollet, G., Dorizzi, B., Les Jardins, J. L., Lunter, J., Ni, Y., and Petrovska-Delacrétaz, D. (2003). Biomet: a multimodal person authentication database including face, voice, fingerprint, hand and signature modalities. In *International Conference on Audio-and Video-based Biometric Person Authentication*, pages 845–853. Springer.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 369 of 404

Impedovo, D. and Pirlo, G. (2008). Automatic signature verification: the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(5):609–635.

Jain, A. K., Ross, A., and Prabhakar, S. (2004). An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology*, 14(1):4–20.

Kalera, M. K., Srihari, S., and Xu, A. (2004). Offline signature verification and identification using distance statistics. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(07):1339–1360.

Kholmatov, A. and Yanikoglu, B. (2009). SUSIG: an on-line signature database, associated protocols and benchmark results. *Pattern Analysis and Applications*, 12(3):227–236.

Liwicki, M., Malik, M. I., Alewijnse, L., van den Heuvel, E., and Found, B. (2012). ICFHR 2012 competition on automatic forensic signature verification (4NSIGCOMP 2012). In *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*, pages 823–828. IEEE.

Liwicki, M., Malik, M. I., van den Heuvel, C. E., Chen, X., Berger, C., Stoel, R., Blumenstein, M., and Found, B. (2011). Signature verification competition for online and offline skilled forgeries (SigComp2011). In *2011 International Conference on Document Analysis and Recognition*, pages 1480–1484. IEEE.

Liwicki, M., van den Heuvel, C. E., Found, B., and Malik, M. I. (2010). Forensic signature verification competition 4NSigComp2010-detection of simulated and disguised signatures. In *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, pages 715–720. IEEE.

Malik, M. I., Ahmed, S., Dengel, A., and Liwicki, M. (2012). A signature verification framework for digital pen applications. In *Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on*, pages 419–423. IEEE.

Malik, M. I., Ahmed, S., Marcelli, A., Pal, U., Blumenstein, M., Alewijns, L., and Liwicki, M. (2015). ICDAR2015 competition on signature verification and writer identification for on-and off-line skilled forgeries (SigWIcomp2015). In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 1186–1190. IEEE.

Malik, M. I., Liwicki, M., Alewijnse, L., Ohyama, W., Blumenstein, M., and Found, B. (2013). ICDAR 2013 Competitions on Signature Verification and Writer Identification for On- and Offline Skilled Forgeries (SigWiComp 2013). In *2013 12th International Conference on Document Analysis and Recognition*, pages 1477–1483. IEEE.

Martinez-Diaz, M., Fierrez, J., Krish, R. P., and Galbally, J. (2014). Mobile signature verification: feature robustness and performance comparison. *IET Biometrics*, 3(4):267–277.

Nel, E.-M., Du Preez, J. A., and Herbst, B. M. (2005). Estimating the pen trajectories of static signatures using Hidden Markov models. *IEEE transactions on pattern analysis and machine intelligence*, 27(11):1733–1746.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 370 of 404

Ortega-Garcia, J., Fierrez-Aguilar, J., Simon, D., Gonzalez, J., Faundez-Zanuy, M., Espinosa, V., Satue, A., Hernaez, I., Igarza, J.-J., Vivaracho, C., et al. (2003). MCYT baseline corpus: a bimodal biometric database. *IEE Proceedings-Vision, Image and Signal Processing*, 150(6):395–401.

Pirlo, G., Rizzi, F., Vacca, A., and Impedovo, D. (2015). Interoperability of biometric systems: Analysis of geometric characteristics of handwritten signatures. In *International Conference on Image Analysis and Processing*, pages 242–249. Springer.

Qiao, Y., Liu, J., and Tang, X. (2007). Offline signature verification using online handwriting registration. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

Soleimani, A., Fouladi, K., and Araabi, B. N. (2016). UTSig: A Persian Offline Signature Dataset. *arXiv preprint arXiv:1603.03235*.

Uppalapati, D. (2007). Integration of Offline and Online Signature Verification systems. *Department of Computer Science and Engineering, IIT, Kanpur*.

van den Heuvel, C., Franke, K., and Vuurpijl, L. (2009). The ICDAR 2009 signature verification competition. *ICDAR2009 proceedings*.

Vargas, J. F., Ferrer, M. A., Travieso, C. M., and Alonso, J. B. (2007). Off-line handwritten signature GPDS-960 corpus. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*.

Viard-Gaudin, C., Lallican, P. M., Knerr, S., and Binter, P. (1999). The ireste on/off (ironoff) dual handwriting database. In *Document Analysis and Recognition, 1999. ICDAR'99. Proceedings of the Fifth International Conference on*, pages 455–458. IEEE.

Yeung, D.-Y., Chang, H., Xiong, Y., George, S., Kashi, R., Matsumoto, T., and Rigoll, G. (2004). SVC2004: First international signature verification competition. In *Biometric Authentication*, pages 16–22. Springer.

Zareen, F. J. and Jabin, S. (2016). Authentic mobile-biometric signature verification system. *IET Biometrics*, 5(1):13–19.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 371 of 404

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

In: Handwriting: Recognition, Development and Analysis ISBN: 978-1-53611-937-4
Editors: Byron L. D. Bezerra et al. © 2017 Nova Science Publishers, Inc.

Chapter 15

PROCESSING OF HANDWRITTEN ONLINE SIGNATURES: AN OVERVIEW AND FUTURE TRENDS

Alessandro Balestrucci, Donato Impedovo and Giuseppe Pirlo
Dipartimento di Informatica,
Università degli Studi di Bari, Italy

1. Introduction

In the modern society, along with the growing need for secure personal recognition, there is an increasing interest toward biometric systems. In fact, while traditional techniques perform personal recognition considering the possession of a token or the knowledge of something, biometric techniques use physiological or behavioural traits of the individual. Physiological traits are based on the measurement of physical characteristics of users, like fingerprint, retina and iris, hand-geometry, face. Behavioural traits are related to behavioural characteristics of users, like voice, keystroke dynamics, handwritten signature (Boyer et al., 2007; Phillips et al., 2000).

Today, in the era of the networked society, handwritten signature is rightly considered a very special biometric trait. The verification of a person's identity by signature analysis does not involve an invasive measurement procedure and people are habituate to using the handwritten signatures as a means for personal verification in their daily lives. In addition, handwritten signatures are a long been established means of personal identification, and their use well-recognized by administrative and financial institutions (Plamondon and Srihari, 2000; Vielhauer, 2005). On the basis of the data acquisition method, two categories of systems for handwritten signature verification can be identified: static (off-line) systems and dynamic (on-line) systems. Static systems use off line acquisition devices which perform data acquisition after the writing process has been completed. Dynamic systems use on line acquisition devices that generate electronic signals representative of the signature during the writing process (Impedovo and Pirlo, 2008; Plamondon and Lorette, 1989; Plamondon, 1994). Of course, integration of low-cost systems for on-line handwriting acquisition in a multitude of personal devices, like tablets, smartphones and PDAs, makes on-line signature verification a relevant opportunity for a multitude of daily activities (Plamondon et al., 2014). The interest toward online signatures is also demonstrated by the standardization of

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

signature data interchange formats that has been supported by several national associations and international bodies, as well as the definition of specific regulations on signature-based personal identity verification systems and procedures (ANSI, 2005; ISO, 2007).

Therefore, it is not surprising the efforts that in the last years more and more researchers from academies and companies have devoted to the field. In fact, automatic signature verification involves aspects from disciplines ranging from neuroscience to engineering, from computer science to human anatomy (Impedovo and Pirlo, 2008), since a handwritten signature is the result of a complex process originating in the signer's brain as a motor control "program", and implemented through the neuromuscular system (Plamondon, 1995; Plamondon and Djioua, 2006).

This chapter provides an overview on the field of on-line signature verification. The chapter is organised as follows: Section "Signature Verification System" describes the structure of a generic system for automatic signature verification. Section "Data Acquisition and Preprocessing" deals with the data acquisition process. The feature extraction phase is discussed in Section "Feature Extraction". Section "Verification" presents the main classification techniques used in the field of automatic signature verification. In Section "On-line Signature Verification Systems" a comparative analysis of the state-of-the-art systems for on-line signature verification is presented. Section "Discussion and Conclusion" addresses the conclusion of the chapter, and provides some considerations of the most valuable directions for future research in the field.

2. Signature Verification System

The process of automatic signature verification can be decomposed into three main phases: data acquisition and pre-processing, feature extraction, verification (Impedovo and Pirlo, 2008; Plamondon and Lorette, 1989).

When on-line signatures are considered, the acquisition device produces electronic signals representative of the signature during the writing process. In this case, a spatio-temporal representation of the signature is generated and the signature is represented as a sequence $\{S(n)\}_{n=0,1...N}$, where $S(n)$ is the signal value sampled at time $n{\cdot}\Delta t$ of the signing process $(0{\leq}n{\leq}N)$, Δt being the sampling period. In the preprocessing phase, the enhancement of the input data is generally based on techniques originating from standard signal processing algorithms, like filtering, noise reduction smoothing and signature normalization in the domain of position, size, orientation and time duration (Impedovo and Pirlo, 2008).

Feature extraction is a very critical phase in the field of automatic signature verification. Two types of features can be considered: functions or parameters. When function features are used, the signature is characterized by a time function, the values of which constitute the feature set. When parameter features are used, the signature is characterized as a vector of parameters, each of which represents the value of one feature (Plamondon and Lorette, 1989).

In the verification process, the authenticity of the test signature is evaluated by matching its features against those stored in the knowledge base developed during the enrolment stage. Two different matching approaches can be considered: distance-based and model-based. Distance-based approaches provide the verification response on the basis of the value

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 374 of 404

of the distance between the test signature and one or more reference signatures. Model-based approaches verify the test signature by estimating how much it fits the signature reference model of the user (Impedovo and Pirlo, 2008; Plamondon and Lorette, 1989).

3. Data Acquisition and Preprocessing

Current technology makes available a lot of devices for data acquisition, capable of detecting position, velocity, acceleration, pressure, pen inclination and forces of the writing process. Traditional digitizing tablets have been overcome by modern electronic pens with digital-ink technology that are easy to integrate into current systems and provide immediate feedback to the writer (Plimmer et al., 2006). More recently, some input devices use ink pen that also produces an exact electronic replica of the actual handwriting. The advantage is the possibility to record on-line and off-line data at the same time (Garcia-Salicetti et al., 2003; Ortega-Garcia et al., 2003b, 2010).

More and more mobile devices are also available for on-line handwriting acquisition, like tablets, smartphones and PDA (Alonso-Fernandez et al., 2005; Elliott, 2004; Plimmer et al., 2006). A special stylus conveying a small CCD camera was developed that captures a series of snapshots of the writing (Nabeshima et al., 1995). The system recovers the whole handwritten trace by analysing the sequence of successive snapshots. There are also approaches which exploit a video camera that is focused on the user while writing on a piece of paper with a normal pen (Bunke et al., 1999). In this way, handwriting is recovered from the spatio-temporal representation given by the sequence of images (Munich and Perona, 2002, 2003). In addition, a hand-glove device for virtual reality applications has been also considered for on-line signature verification (Kamel et al., 2008; Tolba, 1999).

Of course, many interoperability problems arise when signatures are acquired using different acquisition devices. Furthermore, signature verification on handheld devices poses new specific issues concerning quality of signature specimens. In fact, a handwritten signature is generally considered as a well-learned movement. This hypothesis might not be completely appreciated when small and mobile devices are used, along with not standard posture of the signer (i.e. a moving signer) (Tolosana et al., 2015).

Preprocessing generally involves standard filtering and noise reduction, as well as position, size and time duration normalization. Signature segmentation is a task that strongly determines the elements of the signature model and influences all the successive phases of signature verification (Impedovo and Pirlo, 2008).

Some segmentation techniques consider a signature as a sequence of writing units, the "regular" parts of the signature, delimited by abrupt interruptions, the "singularities" of the signature (Impedovo and Pirlo, 2008; Plamondon, 1995). The signature can also be segmented according to the perceptually important points. The importance of a point depends on the extent of change in the writing angle between the selected point and its neighbours. End-points of pen-down strokes can also be considered as significant splitting points (Brault and Plamondon, 1993b). Other approaches use perceptually important points for segmenting signatures, while considering the evolutionary distance measure, based on arc-length distance, for segment association (Schomaker and Plamondon, 1990).

In some cases, segmentation techniques are based on splitting strategies supporting specific signature verification approaches. For example, with Dynamic Time Warping (DTW),

Attorney Docket No. 16-CV-00337-WHO
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 375 of 404

two or more signatures can be segmented into the same number of segments that correspond more or less perfectly (Dimauro et al., 1993; Lee et al., 2004; Yue and Wijesoma, 2000). A model-guided segmentation technique has also been proposed, which uses DTW to segment an input signature based on its correspondence to the reference model (Impedovo and Pirlo, 2008).

4. Feature Extraction

Two types of feature can be considered for signature verification: functions or parameters. When function features are used, the signature is characterized by a time function, the values of which constitute the feature set. When parameter features are used, the signature is characterized as a vector of parameters, each of which represents the value of one feature (Plamondon and Lorette, 1989).

Typical function features are the position, velocity and acceleration functions, as well as pressure, force and direction of pen movements. The position function is conveyed directly by the acquisition device. Conversely, velocity and acceleration functions can be derived numerically from position (Di Lecce et al., 1999; Wu et al., 1997b). Specific devices are now available to capture pressure and force functions directly during the signing process (Wang et al., 2010). Recently, pressure function was used (Hongo et al., 2005; Huang and Yan, 2003; KOMIYA et al., 2001; Ortega-Garcia et al., 2003a; Qu et al., 2003; Schmidt and Kraiss, 1997). Directions of the pen movements and the pen inclinations have been also considered for automatic signature verification (Igarza et al., 2003; Ortega-Garcia et al., 2003a).

Concerning consistency of features for on-line signature verification, a comparative study has demonstrated that the position, velocity and pen inclination functions can be considered to be among the most consistent when a distance-based consistency model is applied (Lei and Govindaraju, 2005). Robustness analysis of function features by personal entropy, to short-term and long-term variability, demonstrated that position is the most robust feature (Houmani et al., 2008).

Typical parameter features are the total signature duration (Kashi et al., 1998; Lee et al., 1996; Qu et al., 2003), the number of pen lifts (pen-down, pen-up) (Lee et al., 1996; Qu et al., 2003), the pen-down/pen-down time ratio (Kashi et al., 1998; Nelson et al., 1994), and parameters derived from the analysis of direction, curvature and moments of the signature trace. Another wide set of parameters can be derived numerically from the representative time functions of a signature: the average (AVE), root mean square (RMS), maximum (MAX) and minimum (MIN) values of position, displacement, speed and acceleration (Lee et al., 1996; Nelson et al., 1994; Qu et al., 2003). Also correlations and time-dependent relations between function features can be considered as parameter features (Kashi et al., 1998; Lee et al., 1996; Nelson et al., 1994) as well as coefficients derived from Fourier- (Dimauro et al., 1994; Wu et al., 1997a, 1998) and Wavelet (Lejtman and George, 2001; Ortega-Garcia et al., 2003a; Nakanishi et al., 2004) transforms.

Some major function and parameter features for on-line signature verification are listed in Table 1.

Of course, features can be considered at global or local level. At global level features reflect the holistic characteristics of the signature. At local level, features describe some

Exhibit 1014 - Page 376 of 404
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

Table 1. Feature Types

Function Features	
Position	(Hongo et al., 2005), (KOMIYA et al., 2001), (Ortega-Garcia et al., 2003a), (Wu et al., 1997b).
Velocity	(Di Lecce et al., 2000, 1999), (Fuentes et al., 2002), (Huang and Yan, 1995), (Jain et al., 2002), (Ortega-Garcia et al., 2003a), (Qu et al., 2003), (Schmidt and Kraiss, 1997), (Wu et al., 1997a).
Acceleration	(Schmidt and Kraiss, 1997).
Pressure	(Hongo et al., 2005), (Huang and Yan, 1995), (KOMIYA et al., 2001), (Ortega-Garcia et al., 2003a), (Qu et al., 2003), (Schmidt and Kraiss, 1997).
Direction of pen movement	(Fuentes et al., 2002), (Igarza et al., 2003).
Pen inclination	(Igarza et al., 2003), (KOMIYA et al., 2001), (Martens and Claesen, 1998), (Ortega-Garcia et al., 2003a).
Parameter Features	
Total signature duration	(Kashi et al., 1998), (Lee et al., 2004), (Lee et al., 1996), (Nelson et al., 1994), (Qu et al., 2003).
Pen-down time ratio	(Kashi et al., 1998), (Nelson et al., 1994).
Number of Pen-Ups/Pen-Downs	(Lee et al., 2004), (Lee et al., 1996), (Qu et al., 2003).
Direction-based	(Kashi et al., 1998), (Lee et al., 2004), (Nelson et al., 1994), (Qu et al., 2003), (Zou et al., 2003).
Curvature-based	(Jain et al., 2002).
Moment-based	(Kashi et al., 1998).
AVE/ RMS/ MAX/ MIN of Posit., Displ., Speed, Accel.	(Fuentes et al., 2002), (Kashi et al., 1998), (Khan et al., 2006), (Lee et al., 2004), (Lee et al., 1996), (Nelson et al., 1994), (Qu et al., 2003).
Duration of Positive/Negative Posit., Displ., Speed, Accel.	(Kashi et al., 1998), (Lee et al., 2004), (Lee et al., 1996), (Nelson et al., 1994).
X-Y correlation of Posit., Displ., Speed, Accel.	(Fuentes et al., 2002), (Kashi et al., 1998), (Nelson et al., 1994).
Fourier Transform	(Dimauro et al., 1994), (Wu et al., 1997a, 1998).
Wavelet Transform	(Lejtman and George, 2001), (Martens and Claesen, 1998), (Nakanishi et al., 2004).

very specific characteristics of a signature region. For instance, function features can be considered at global level, i.e. for the whole signature, or at local level, i.e. for individual signature segments. Concerning parameters, typical global parameters are total duration, number of pen lifts and number of components of the signature. Typical local parameters are related to direction-based, curvature-based and moment-based features estimated at regional level of a signature (Impedovo and Pirlo, 2008).

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 377 of 404

5. Verification

The aim of the verification phase is to evaluate the authenticity of a test signature by matching its features against those stored in the knowledge base developed during the enrolment stage. The result for the authenticity of a signature can be provided as a Boolean value or a real value, when a confidence value concerning the decision is required. In the literature, two main types of comparison techniques were considered: distance-based and model-based techniques (Impedovo and Pirlo, 2008; Plamondon and Lorette, 1989).

When function features are considered, Dynamic Time Warping (DTW) was the most exploited technique for signature comparison, since this allows the time axis of two time sequences that represent a signature to be compressed or expanded locally, in order to obtain the minimum of a given distance value (Impedovo and Pirlo, 2008; Parizeau and Plamondon, 1990). In order to reduce the computational cost of the comparison, several advanced DTW strategies were proposed for data reduction based on Genetic Algorithms (GA), Principal or Minor Component Analysis (PCA/MCA) and Linear Regression (LR) (Impedovo and Pirlo, 2008). When parameters are considered as features, both Euclidean (Khan et al., 2006) and Mahalanobis (Martens and Claesen, 1998; Nyssen et al., 2002) distance have been considered for matching the target specimen to the parameter-based model of the declared signer's signatures. Similarity measures (Wu et al., 1997a), split-and-merge strategies (Wu et al., 1997b) and string-matching (Chen and Ding, 2002) have also been considered for signature comparison. Another effective approach for on-line signature verification uses Support Vector Machines (SVM), that can map input vectors to a higher dimensional space, in which clusters may be determined by a maximal separation hyper-plane (Fuentes et al., 2002; Kholmatov and Yanikoglu, 2005).

Some of the most valuable model-based techniques for signature comparison concern neural networks, multi-layer perceptrons (MLP), time-delay neural networks, backpropagation networks, self-organizing map (Fuentes et al., 2002; Huang and Yan, 1995; Lejtman and George, 2001; Wessels and Omlin, 2000). Of course, the use of neural network for on-line signature verification can be limited by the fact they generally require large amounts of learning data, which are not available in many applications (Impedovo and Pirlo, 2008; Leclerc and Plamondon, 1994). Another model-based comparison technique very effective for signature modelling uses Hidden Markov Models (HMM). Studies have found that HMM is highly adaptable to personal variability (Fuentes et al., 2002; Martínez Díaz et al., 2008; Van et al., 2007), therefore they can support effective signature modelling techniques (Huang and Yan, 2003). Although both Left-to-Right (LR) and Ergodic (E) topologies have been considered In the literature for on-line signature verification, the LR topology is generally considered best suited to signature modelling (Igarza et al., 2003; Woch et al., 2011; Zou et al., 2003).

Table 2 shows some of the most valuable distance-based and model-based techniques for signature comparison.

Multi-expert approaches based on abstract-level, ranked-level and measurement-level combination methods have also be considered in order to improve verification performance (Hongo et al., 2005; Nanni et al., 2010; Nanni, 2006). In particular, multi-expert approaches have been used for implementing top-down and bottom-up signature verification strategies (Impedovo and Pirlo, 2007; Pirlo, 1994). When bottom-up strategies are considered, a sig-

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 378 of 404

Table 2. Comparison Techniques

Distance - based technique		
Dynamic Programming		(Lee et al., 2004), (Nakanishi et al., 2004)
Dynamic Time Warping (DTW)	Continuous	(Bovino et al., 2003), (Di Lecce et al., 2000, 1999), (Dimauro et al., 1993, 2002), (Huang and Yan, 2003), (Munich and Perona, 2003).
	GA-based	(Wirotius et al., 2005)
	PCA-based	(Kholmatov and Yanikoglu, 2005), (Li et al., 2004).
	MCA-based	(Li et al., 2004).
	PA-based	(Wirotius et al., 2005).
	EP-based	(Feng and Wah, 2003).
	Random-based	(Wirotius et al., 2005).
	Asymmetric	(Martens and Claesen, 1998).
Correlation		(Parizeau and Plamondon, 1990).
Split-and-Merge		(Wu et al., 1997b).
String- / Graph- / Tree-matching		(Chen and Ding, 2002), (Jain et al., 2002), (Parizeau and Plamondon, 1990).
Structural Description Graph		(Dimauro et al., 1994), (Nyssen et al., 2002).
Euclidean Distance		(Khan et al., 2006).
Mahalanobis Distance		(Martens and Claesen, 1998), (Nyssen et al., 2002).
Membership functions		(Qu et al., 2003).
Dynamic Similarity Measure		(Wu et al., 1998).
Support Vector Machine (SVM)		(Fuentes et al., 2002), (Ibrahim et al., 2010), (Kholmatov and Yanikoglu, 2005).
Model - based technique		
Hidden Markov Models (HMM)	Left-to-Right topology	(Dolfing et al., 1998), (Fuentes et al., 2002), (Galbally et al., 2009b,a), (Igarza et al., 2003, 2004), (Kashi et al., 1998), (Martínez Díaz et al., 2008), (Ortega-Garcia et al., 2003a), (Shafiei and Rabiee, 2003), (Van et al., 2007), (Wessels and Omlin, 2000), (Yoon et al., 2002), (Zou et al., 2003).
	Ergodic topology	(Wessels and Omlin, 2000).
Neural Network (NN)	Multi-layer perceptrons (MLP)	(Fuentes et al., 2002), (Huang and Yan, 1995).
	Backpropagation Network (BPN)	(Lejtman and George, 2001).
	Self-organizing Map	(Wessels and Omlin, 2000).

nature is verified starting with the analysis of its basic elements, like strokes or components. This approach can lead to lower error rates, compared to global approaches, since a large amount of personal information is conveyed in specific parts of the signature and cannot be detected when the signature is viewed as a whole (Brault and Plamondon, 1993b; Dimauro et al., 1993, 1994; Schmidt and Kraiss, 1997). When top-down verification strategies are used, hybrid topologies have shown to combine the performance advantages of serial approaches in quickly rejecting very poor forgeries, while retaining the reliability of parallel combination schemes. For example, multi-level verification systems first verify the structural organization of a target signature and then analyse in detail the authenticity of its individual elements (Dimauro et al., 1994; Huang and Yan, 2003).

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 379 of 404

6. On-line Signature Verification Systems

In recent years the need to move the development of on-line signature verification systems from academic research to on field applications makes the problem of performance analysis more and more important.

Several standard measures are generally considered for evaluating the performance of on-line signature verification systems (Leclerc and Plamondon, 1994; Plamondon and Lorette, 1989): (1) the Type I and II error rates, which concern the false rejection of genuine signatures (FRR-False Rejection Rate) and the false acceptance of forged signatures (FAR-False Acceptance Rate), respectively; (2) the Equal Error Rate (EER) that is a measure of the overall error of a system. It is defined as the system error rate when FRR=FAR; (3) the Receiver Operating Characteristic (ROC) curve, that plots the FRR vs. the FAR. The ROC curve has several useful properties: the Area Under the Curve (AUC) of the ROC can be used to estimate system performance by using a single value since the AUC provides the probability that the classifier will rank a randomly chosen positive sample higher than a randomly chosen negative sample.

Of course, performance evaluation is a very critical and complex task (Plamondon and Lorette, 1989; Plamondon and Srihari, 2000): the existence of skilled forgeries for a given signature is not certain, nor is the possibility of collecting good quality forgery samples for the test (Alonso-Fernandez et al., 2009; Ballard et al., 2007). For the purpose different classes of forgeries and several public signature databases for experimental tests have been considered. Three different classes of forgeries are generally used (Plamondon and Lorette, 1989): (a) random forgeries, in which the forger uses his own signature instead of the signature to be tested; (b) simple forgeries, in which the forger makes no attempt to simulate or trace a genuine signature; (c) freehand or skilled forgeries, in which the forger tries and practices imitating as closely as possible the static and dynamic information of a genuine signature. Moreover, although the development of a benchmark signature database is a time-consuming and expensive process, several public signature databases were realized for supporting the crucial step of the comparative assessment of the approaches proposed in the literature (Ortega-Garcia et al., 2010; Phillips et al., 2000). Among others, some of the most important benchmark databases are the Philips database (Dolfing et al., 1998; Phillips et al., 2000), the BIOMET multimodal database (Garcia-Salicetti et al., 2003), the MCYT database (Ortega-Garcia et al., 2003b), the SVC2004 database (Yeung et al., 2004), the Biosecure multimodal database (Ortega-Garcia et al., 2010) the Caltech database (that contains signatures acquired by a camera) (Munich and Perona, 2003). More recently, two well-defined strategies for the generation of synthetic signatures have been considered, in order to overcome the problem of lacking of on-line signatures (Galbally et al., 2012b,a). The first strategy concerns the synthetic sample generation, in which synthetic samples are obtained from the real ones of a given individual (Fuentes et al., 2002; Rabasse et al., 2007). The second strategy concerns the synthetic individual generation, in which synthetic samples are generated according to a model of the signature produced by a population of individuals (Fuentes et al., 2002; Galbally et al., 2010)

In the following, the performance of some of the most valuable on-line signature verification systems in the literature are discussed.

Table 3 lists some on-line signature verification systems using distance-based classifica-

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 380 of 404

tion techniques. Arora et al. (2014) used both the discrete fractional cosine transformation (DFrCT) and the discrete cosine transform (DCT) for feature extraction. The experimental results demonstrate the superiority of DFrCT-based features with respect to DCT-based features. DTW was used by Bovino et al. (2003), who presented a multi-expert system based on a stroke-oriented description of signatures. Each stroke was analysed in the position, velocity and acceleration domain. A two-level scheme is used for decision combination. For each stroke, soft- and hard-combination rules were used at the first level to combine decisions obtained by DTW from different representation domains. Simple and weighted averaging was used at the second level to combine decisions from different parts of the signature. Di Lecce et al. (2000) performed signature verification by combining the efforts of three experts. The first expert used shape-based features and performed signature verification by means of global analysis. The second and third experts used speed-based features and a regional analysis. The expert decisions were combined by a Majority Voting strategy. The system proposed by Gomez-Barrero et al. (2015) combined the high performance of DTW-based systems in verification tasks, with the high potential for skilled forgery detection of the features of the Sigma-LogNormal model. Griechisch et al. (2014) used position-based, velocity-based and pressure features and exploited the potential of the Kolmogorov-Smirnov statistics in on-line signature verification. Signature comparison is performed by a method based on the distribution distance determined by applying the Kolmogorov-Smirnov test. Guru and Prakash (2009) represented on-line signatures by interval-valued symbolic features. They used parameter-based features derived by a global analysis of signatures and achieved the best verification results when a writer-dependent threshold was adopted for distance-based classification. Jain et al. (2002) used a set of local parameters which described both spatial and temporal information. In the verification process, string-matching was used to compare the test signature to all the signatures in the reference set. Three methods were investigated to combine the individual dissimilarity values into one value: the minimum, the average and the maximum of all the dissimilarity values. Common and personalized (signer-dependent) thresholds were also considered. The best results were achieved by considering the minimum of all the dissimilarity values combined with personalized threshold values. In the system by Nakanishi et al. (2004), the position signals of an on-line signature are decomposed into sub-band signals using the Discrete Wavelet Transform (DWT), and Dynamic Programming was used for signature matching. The system by Pirlo et al. (2015a) considered four domains of representation of on-line signature: position, velocity, acceleration and pressure. They used stability information to detect the most profitable domain of representation of a signature for verification aims, according to a local analysis strategy (Impedovo and Pirlo, 2010). Successively, local verification decisions obtained by DTW were combined to provide the verification decision of the entire signature. Wibowo et al. (2014) considered stable features of signature that are K-L coefficients of the forward and backward variances between the reference and signature to be verified. Fourier analysis was applied by Wu et al. (1998) for on-line signature verification, as they extracted and used cepstrum coefficients for verification, according to a dynamic similarity measure. Yeung et al. (2004) reported the results of the First International Signature Verification Competition (SVC2004), in which teams from all over the world participated. SVC2004 considered two separate signature verification tasks using two different signature databases. The signature data for the first task contained position

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 381 of 404

information only. The signature data for the second task contained position, pen inclination and pressure information. In both cases, DTW-based approaches provided the best results. The system presented by Nyssen et al. (2002) used global, local and function features. In the first verification stage, a parameter-based method was implemented, in which the Mahalanobis distance was used as a measure of dissimilarity between the signatures. The second verification stage involved corner extraction and corner matching, as well as signature segmentation. In the third verification stage, an elastic matching algorithm was used, and a point-to-point correspondence was established between the compared signatures. By combining the three different types of verification, a high security level was reached. Zou et al. (2003) used local shape analysis for on-line signature verification. Specifically, FTT was used to derive spectral and tremor features from well-defined segments of the signature. A weighted distance was finally considered, in order to combine the similarity values derived from the various feature sets.

Table 3 reports some on-line signature verification systems using model-based classification techniques. Igarza et al. (2003) also used a Left-to-Right HMM for on-line signature verification and demonstrated its superiority over Ergodic HMMs. The superiority of Principal Component Analysis and Minor Component Analysis for on-line signature verification over DTW and Euclidean-based verification was also investigated and demonstrated by Igarza et al. (2004). The on-line signature verification system proposed by Kashi et al. (1997) used a Fourier transform-based normalization technique, and both global and local features for signature modelling. The global features captured the spatial and temporal information of the signature, and the local features, extracted by a Left-to-Right HMM, captured the dynamics of the signature production process. The verification result was achieved by combining the information derived by both global and local features. Lee et al. (2004) performed signature verification by means of a back-propagation neural network, which integrated verification results at segment level, using a bottom-up strategy. They performed signature segmentation by means of a DP technique based on geometric extrema. Segment matching was performed by global features and a DP approach. Martínez Díaz et al. (2008) presented a Left-to-Right HMM-based signature verification system for handheld devices. Signatures were captured by a PDA and described by the position function only. The best results were achieved by user-dependent HMM modelling. Muramatsu and Matsumoto (2003) used HMM-LR to incorporate signature trajectories for on-line signature verification. Individual features were extracted as high frequency signals in sub-bands. The total decision for verification was carried out by averaging the verification results achieved at each sub-band. Ortega-Garcia et al. (2003a) presented an investigation of the ability of HMM-LR to model the signing process, based on a set of 24 function features (8 basic function features and their first and second derivatives). In Shafiei and Rabiee's system (Shafiei and Rabiee, 2003), each signature was segmented using its perceptually important points. For every segment, four dynamic and three static parameters were extracted, which are scale- and displacement-invariant. HMM was used for signature verification. Swanepoel and Coetzer (2014) used pen-positions, pressure and pen-tilt as features and adopted a dynamic time warping-based dichotomy transformation and a writer-specific dissimilarity normalisation technique for signature representation. Signature comparison is performed by a support vector machines with linear and radial basis function kernels. The results demonstrate that non-linear kernel significantly outperforms linear kernel. Wessels

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 382 of 404

Table 3. System Performances: distance-based techniques

Full Database (FD) , Signature (S), Genuine Signatures (G) , Forgeries (F), Random Forgeries (RF), Simple Forgeries (SF), Skilled Forgeries (SK), Number of Authors (A)

Matching Technique	Main features	Database	Results	Reference
Euclidean Distance	**Test 1:** DFrCT **Test 2:** CDT	FD (SVC 2004) 100(G)(20(G)x5(A)) 100(F)(20(F)x5(A))	**(Test 1) EER:** 5% **(Test 2) EER:** 7.04%	(Arora et al., 2014)
DTW (ME by simple averaging)	Position, velocity, acceleration	**Training)** 45(G) (3(G) x 15(A)) **Test)** 750(G) (50(G)x15(A)) , 750 (F) (50(F)x15(A))	**EER:** 0,4%	(Bovino et al., 2003)
DTW (ME by majority voting)	Shape-based features (segmentation-dependent), Velocity	**Training)** 45(G) (3(G) x 15(A)) **Test)** 750 (G) (50(G)x15(A)) , 750 (F) (50(F)x15(A))	**FRR:** 3,2% . **FAR:** 0.55%	(Di Lecce et al., 2000)
DTW	Log-Normal-based Features	**Training Set)** 16 (G) x 50(A), 12 (SF)x50(A) **Development set)** 16 (G) x 100(A), 12 (SF)x100(A) **Test set)** 16 (G) x 250(A), 12 (SF)x250(A)	**Test 1)** DTW EER: 5.80% (SF), 1.07% (RF) **Test 2)** DTW + SL. EER: 4.77 (SF), 0.50 (RF)	(Gomez-Barrero et al., 2015)
Kolmogorov-Smirnov Distance.	position-based, velocity-based, pressure	**FD)** (12(G x10(A), 24(F)x10(A))	**EER:** 13%	(Griechisch et al., 2014)
Distance-based	Total signature duration, number of pen-ups, STD velocity and acceleration in x and y direction, number of local maxima in x direction, etc.	**FD1)** **Training1)** 2000 (G) (20(G) x 100(A)) **Test1)** 500 (G) (5(G) x 100 (A)), 9900 (RF) (99 (RF) x 100(A)) , 500 (SK) (5(SK) x 100(A)) **FD2)** **Training2)** 6600(G) (20(G) x 330(A)) **Test2)** 1650 (G) (5(G) x 330(A)), 75570 (RF) (229(RF) x 330(A)) , 8250(SK) (25(SK) x 330(A))	**FD1)** **EER :** 3.80% (Test1 - SK) **EER :** 1.65% (Test1 - RF) **FD2)** **EER :** 4.70% (Test2 - SK) **EER :** 1.67% (Test2 - RF)	(Guru and Prakash, 2009)
DTW	Velocity, pressure	**FD)** 4600 (S)	**EER:** 4%	(Huang and Yan, 1995)
String-matching	Velocity, curvature-based.	**FD)** 1232 (S) (from 102 (A))	**FRR:** 3,3%.. **FAR:** 2,7% (common threshoold) **FRR:** 2,8%. **FAR:** 1,6% (writer dependent threshold)	(Jain et al., 2002)
DTW (PCA, MCA)	Position , velocity	**Training)** 405 (G) (5(G) x 81(A)) **Test)** 405 (G) (5(G) x 81(A)), 405(F) (5(F) x 81(A))	**EER:** 5,00%	(Li et al., 2004)
Dynamic Programming	Wavelet Transform.	**Training)** 20(G) from(4(A)) **Test)** 98 (G) (from 4(A)) , 200(F) (from 5(A))	**EER:** 4%	(Nakanishi et al., 2004)
DTW	Position, velocity, acceleration, pressure	**Training Database)** 15(G)x100(A) 15(SF)x100(A) **Test Database)** 15(G)x100(A) 5(G)x100(A)	**FRR:** 2.15% **FAR:** 2.10%	(Pirlo et al., 2015a)
Euclidean Norm	K-L coefficients	**FD)** 5000 (S): 2500(G) (25(G)x100(A))+2500(F) (25(F)x100(A)) (signature are from the MCYT-100 - 5 signature are considered as reference for each signer)	**EER:** 4.49%	(Wibowo et al., 2014)
Membership function	Total signature time, AVE/RMS speed, pressure, direction-based, number of pen-ups/pen-downs, etc.	**Test)** 60 (G) , 60 (F)	**FRR:** 6,67%. **FAR:** 1,67%	(Qu et al., 2003)
Dynamic similarity measure	Fourier transform (cepstrum coefficients)	**Training)** 270(G) (from 27(A)) **Test)** 560 (G) (from 27(A)), 650 (F)	**FRR:** 1,4% . **FAR:** 2,8%	(Wu et al., 1997a)
DTW (best result)	**Task 1:** Position **Task 2:** Position, pen inclination (azimuth), pressure, etc.	**Training)** 800(G)(20(G)x40(A)),800(SK)(20(SK)x40(A)) **Test 1)** 600(G)(10(G)x60(A)), 1200(SK)(20(SK)x60(A)) **Test 2)** 600(G)(10(G)x60(A)), 1200(RF)(20(RF)x60(A))	**(Test 1) EER:** 2,84% (Task 1). **EER:** 2,89% (Task 2) **(Test 2) EER:** 2,79% (Task 1). **EER:** 2,51% (Task 2)	(Yeung et al., 2004) (1st Signature Verification Competition)
Euclidean Distance, Mahalanobis distance, DTW	Geometric-based, Curvature-based	**FD)** 306 (G) , 302 (F)	**FRR:** 5,8% . **FAR:** 0%	(Nyssen et al., 2002)
Membership function	Speed, pressure, direction-based, Fourier transform	**FD)** 1000 (G) , 10000 (F)	**FRR:** 11,30%. **FAR:** 2,00%	(Zou et al., 2003)

and Omlin (2000) combined a Kohonen self-organizing feature map and HMM. Both Left-to-Right and Ergodic models were considered. Yang et al. (1995) used directional features

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 383 of 404

along with several HMM topologies for signature modelling. The results demonstrated that HMM-LR is superior to other topologies in capturing the individual features of the signatures, while at the same time accepting variability in signing. A polar coordinate system was considered for signature representation by Yoon et al. (2002), whose aim was to reduce normalization error and computing time. Signature modelling and verification were performed by HMMs, which demonstrated their ability to capture the local characteristics in the time sequence data, as well as their flexibility in terms of modelling signature variability.

Table 4. System Performances: model-based techniques

Full Database (FD) , Signature (S), Genuine Signatures (G) , Forgeries (F), Random Forgeries (RF), Simple Forgeries (SF), Skilled Forgeries (SK), Number of Authors (A)

Matching Technique	Main features	Database	Results	Reference
HMM	Direction of pen movement, etc.	**FD)** 3750 (G) (25(G)x150(A)) , 3750 (F)	**EER:** 9,253%	(Igarza et al., 2003)
HMM	Total signature time duration, X-Y (speed) correlation, RMS speed, Moment-based, direction-based, etc.	**Test)** 542 (G) , 325 (F)	**EER:** 2,5%	(Kashi et al., 1997)
NN (with DP)	Position (geometric extrema), AVE velocity, number of pen-ups, time duration of neg. /pos. velocity, total signing time, direction-based, ...	**FD)** 6790 (S) (from 271(A))	**EER:** 0,98%	(Lee et al., 2004)
HMM	Position	**FD)** 1000(G) (20(G) x 50(A)) , 1000(SK) (20(SK) x 50(A)) , **Training)** 250(G) (5(G) x 50(A)) **Test)** 750(G) (15(G)x50(A)) , 2450 (RF) (49(RF) x 50(A)) , 1000 (SK) (20(SK) x 50(A)) ,	**EER:** 5.2% (with **RF**) **EER:** 15.8% (with **SF**)	(Martínez Díaz et al., 2008)
HMM	Direction of pen movements	**Training)** 165 (G) **Test)** 1683 (G) , 3170 (SK)	**EER:** 2,78%	(Muramatsu and Matsumoto, 2003)
HMM	Position, Velocity, Pressure, Pen Inclination (Azimuth), Direction of Pen Movement, ...	**Training)** 300(G) (from 50(A)) **Test)** (450 (G) (from 50(A)), 750 (SK) (from 50(A))	**EER:** 1,21% (global threshold) **EER:** 0,35% (user-specific threshold)	(Ortega-Garcia et al., 2003a)
HMM	AVE Speed, acceleration, pressure, Direction of Pen movement,...	**FD)** 622 (G) (from 69(A)) , 1010 (SK)	**FRR:** 12% . **FAR:** 4%	(Shafiei and Rabiee, 2003)
SVM	horizontal and vertical pen-positions, pressure, pen-tilt in the x and y directions	**FD)** 1530 (G), 3000 (SF)	**Test 1)** EER: 1.26% (with 15 reference signatures per writer) **Test 2)** EER: 3.52% (with 5 reference signatures per writer)	(Swanepoel and Coetzer, 2014)
HMM	Position, Pressure, Direction of pen movements, Pen inclination.	**Training)** 750 (G) (15(G) x 50(A)) **Test)** 750 (G) (15(G) x 50(A))	**FAR:** 13%	(Wessels and Omlin, 2000)
HMM	Direction of Pen movements	**FD)** 496 (S) (from 31 (A)	**FRR:** 1,75% . **FAR:** 4,44%	(Yang et al., 1995)
HMM	Position	**Training)** 1500 (S) ((15 (S) x 100 (A)) **Test)** 500(S) (5 (S) x 100 (A))	**EER:** 2,2%	(Yoon et al., 2002)

Discussion and Conclusion

In our society, along with the increasing security requirements, interest in on-line signature verification continues to grow. Handwritten signature verification is rightly considered a non-invasive and non-threatening process that is compatible with a large variety of daily

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 384 of 404

administrative and financial applications.

This chapter has presented a general description of basic approaches for on-line signature verification and an overview of the most promising systems in the literature. In the near future, in order to make signature verification systems able to operate in daily working scenarios, some specific directions of research need to be furtherly addressed.

First, it is worth noting that although different feature sets and very effective model-based and distance-based matching techniques have been proposed so far, signature verification accuracy still needs to be improved significantly, in order to satisfy some requirements of critical applications. In this direction, two main aspects concern the possibility to use multi-expert approach for signature classification and verification with adaptive capabilities. Multi-expert systems can combine decisions obtained through top-down and bottom-up strategies, according to abstract-level, ranked-level and measurement-level approaches. Furthermore, they can use matching algorithms at both the global and the local level and therefore, they are able to improve significantly verification obtained by an individual classifier. In addition, when multimodal biometrics is considered, the multi-expert approach is useful to integrate decisions provided by uni-modal biometric systems.

Another aspect that need specific attention concerns capability of systems to be adaptive to different signature characteristics. In fact, handwritten signatures can be extremely different patterns, depending on the country, culture, habits, age, physical and physiological characteristics of the signer, etc., hence it is quite obvious that a real-world application of signature verification systems needs to be adaptive enough to handle this huge number of differences. For instance, Western and non-western signatures are very different in style. Western-style signatures generally consist of signs that could form concatenated text combined with pictorial strokes. Conversely, a wide range of styles can be found in non-western signatures. Chinese and Japanese signatures generally consist of independent symbols, Arabic and Persian signatures are cursive sketches that are usually independent of the person's name. Of course, the possibility to find effective solutions depends on the availability of multi-cultural and multi-language benchmark databases of signatures, available to a wide research community, in order to enable comparative evaluation of signature verification systems under different application scenarios. Specific research on these topics are currently in progress (Wolf et al., 2006). In other words, a practical system for signature verification should guarantee the same Quality of Service (QoS), whatever the types and characteristics of the signatures. From the other side, it could be argued that short signatures could convey less information than long signatures, resulting in less accurate verification results. Similarly, correct verification of signatures of people with common name and surname could be more difficult, since signatures could be similar in shape to those of other individuals (Ballard et al., 2007; Parizeau and Plamondon, 1989).

In order to face this kind of problem, specific research have been dedicated to the analysis of signature stability and complexity. Concerning signature stability, a local stability function based on the Dynamic Time Warping (DTW) was proposed (Dimauro et al., 1993; Huang and Yan, 2003). In particular DTW is used to match a genuine signature against other authentic specimens and to identify the Direct Matching Points (DMPs), that are unambiguously matched points of the genuine signature. Hence, a DMP indicates the presence of a small stable region of the signature, since no significant distortion has been locally detected. The local stability of a point of a signature is determined as the average number of

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 385 of 404

time the point is a DMP, when the signature is matched against other genuine signatures. Following this procedure low- and high-stability regions can be identified and used to adapt system behavior to the specific characteristics of the signature (Congedo et al., 1994, 1995; Dimauro et al., 2002). Variability of handwritten patterns have also been estimated using the Kinematic Theory of human movements (Elliott, 2004) or through a client-entropy measure based on local density estimation by a HMM (Salicetti et al., 2008). In particular, the entropy-based measure is used to access whether a signature contains or not enough information to be successfully used for personal verification (Salicetti et al., 2008; Houmani et al., 2008). Stability can be also estimated by the analysis of common features extracted from the signature, in order to obtain global information on signature repeatability (Galbally et al., 2009a; Guest, 2004). From these approaches it results a set of features exists that remain stable over long time periods, while there are other features which change significantly in time (Houmani et al., 2008, 2009; Lei and Govindaraju, 2005). This information can be very useful to improve over time verification performance. In general, stability analysis is used to estimate both short-term and long-term modifications of a signature. Short-term modifications depend on the psychological condition of the writer and on the writing conditions. Information about short-term modification can be used to select the best sub-set of reference signatures and the most effective feature functions for signature verification, while providing useful information to weight the verification decision obtained at the stroke level. Long-term modifications depend on the alteration of the physical writing system of the signer (arm and hand, etc.), as well as on the modification of the motor control "program" in his/her brain (Plamondon et al., 2014). Concerning complexity, it is worth noting that no common meaning of handwriting complexity was defined yet. Notwithstanding, it is generally argued that the complexity of a signature can be critical to the reliability of the examination process (Huber and Headrick, 1999). In general, in signature analysis, signature complexity can be thought to be an estimator of the difficulty for its imitation. Signature complexity can be obtained as the result of the difficulty in perceiving, preparing and executing each stroke of the signature itself (Brault and Plamondon, 1993a). A complexity theory, which is based on the theoretical relationship between the complexity of features of the handwriting process and the number of concatenated strokes, was also considered for complexity estimation. According to this theory signature complexity can be estimated by analyzing variables that are indirectly related to the number of concatenated strokes, like for instance the number of turning points, the number of feathering points, and the number of intersections and retraces (Found and Rogers, 1995).

In addition, it is worth noting that as the number of devices available for signature acquisition is continuously growing, device interoperability is an even more relevant issue that needs specific research. In fact, signature signals can change significantly depending on the type of the acquisition device, writing area and stylus type, but also on the basis of modification of personal characteristics (Guest, 2006; Guest and Fairhurst, 2006; Impedovo and Pirlo, 2008; Maiorana et al., 2010).

Finally, new interesting directions of research have been devoted to developing of handwriting-based cryptosystems (Uludag et al., 2004) as well as to the use of handwritten signatures to discriminate the health conditions of the subject. Recent studies have been devoted to the analysis of handwriting (Djioua and Plamondon, 2009; Longstaff and Heath, 2006; O'Reilly and Plamondon, 2009, 2012a; Woch et al., 2011) but some research are de-

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 386 of 404

voted to the use of signatures to detect brain stroke risk factors (O'Reilly and Plamondon, 2011, 2012b) or to the diagnosis of Parkinson (Rosenblum et al., 2013; Van Gemmert et al., 2003) and Alzheimer (Impedovo et al., 2013; Pirlo et al., 2015b; Yan et al., 2008) diseases.

References

Alonso-Fernandez, F., Fierrez, J., Gilperez, A., Galbally, J., and Ortega-Garcia, J. (2009). Robustness of signature verification systems to imitators with increasing skills. In Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on, pages 728–732. IEEE.

Alonso-Fernandez, F., Fierrez-Aguilar, J., and Ortega-Garcia, J. (2005). Sensor interoperability and fusion in signature verification: A case study using tablet pc. In *Advances in Biometric Person Authentication*, pages 180–187. Springer.

ANSI (2005). Information technology biometric data interchange formats signature/sign data. *ANSI INCITS 395-2005*.

Arora, M., Singh, K., and Mander, G. (2014). Discrete fractional cosine transform based online handwritten signature verification. In *Engineering and Computational Sciences (RAECS), 2014 Recent Advances in*, pages 1–6. IEEE.

Ballard, L., Lopresti, D., and Monrose, F. (2007). Forgery quality and its implications for behavioral biometric security. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 37(5):1107–1118.

Bovino, L., Impedovo, S., Pirlo, G., and Sarcinella, L. (2003). Multi-expert verification of hand-written signatures. In *ICDAR*, volume 3, pages 932–936.

Boyer, K. W., Govindaraju, V., and Ratha, N. K. (2007). Special issue on recent advances in biometric systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 37(5).

Brault, J.-J. and Plamondon, R. (1993a). A complexity measure of handwritten curves: modeling of dynamic signature forgery. *IEEE transactions on Systems, Man, and Cybernetics*, 23(2):400–413.

Brault, J.-J. and Plamondon, R. (1993b). Segmenting handwritten signatures at their perceptually important points. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):953–957.

Bunke, H., Von Siebenthal, T., Yamasaki, T., and Schenkel, M. (1999). Online handwriting data acquisition using a video camera. In Document Analysis and Recognition, 1999. ICDAR'99. Proceedings of the Fifth International Conference on, pages 573–576. IEEE.

Chen, Y. and Ding, X. (2002). On-line signature verification using direction sequence string matching. In Second International Conference on Image and Graphics, pages 744–749. International Society for Optics and Photonics.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 387 of 404

Congedo, G., Dimauro, G., Forte, A., Impedovo, S., and Pirlo, G. (1995). Selecting reference signatures for on-line signature verification. In International Conference on Image Analysis and Processing, pages 521–526. Springer.

Congedo, G., Dimauro, G., Impedovo, S., and Pirlo, G. (1994). A new methodology for the measurement of local stability in dynamical signatures. In 4th International Workshop on Frontiers in Handwriting Recognition, pages 135–144.

Di Lecce, V., Dimauro, G., Guerriero, A., Impedovo, S., Pirlo, G., and Salzo, A. (1999). Image basic features indexing techniques for video skimming. In Image Analysis and Processing, 1999. Proceedings. International Conference on, pages 715–720. IEEE.

Di Lecce, V., Dimauro, G., Guerriero, A., Impedovo, S., Pirlo, G., and Salzo, A. (2000). A multi-expert system for dynamic signature verification. In International Workshop on Multiple Classifier Systems, pages 320–329. Springer.

Dimauro, G., Impedovo, S., Modugno, R., Pirlo, G., and Sarcinella, L. (2002). Analysis of stability in hand-written dynamic signatures. In Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on, pages 259–263. IEEE.

Dimauro, G., Impedovo, S., and Pirlo, G. (1993). A signature verification system based on dynamical segmentation technique. In Proceedings of the International Workshop on Frontiers in Handwriting Recognition, pages 262–271.

Dimauro, G., Impedovo, S., and Pirlo, G. (1994). Component-oriented algorithms for signature verification. *International Journal of Pattern Recognition and Artificial Intelligence*, 8(03):771–793.

Djioua, M. and Plamondon, R. (2009). A new algorithm and system for the characterization of handwriting strokes with delta-lognormal parameters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2060–2072.

Dolfing, J., Aarts, E., and Van Oosterhout, J. (1998). On-line verification signature with hidden markov models. In Proc. 14th International Conference Pattern Recognition, pages 1–309.

Elliott, S. J. (2004). Differentiation of signature traits vis-a-vis mobile-and table-based digitizers. *ETRI journal*, 26(6):641–646.

Feng, H. and Wah, C. C. (2003). Online signature verification using a new extreme points warping technique. *Pattern Recognition Letters*, 24(16):2943–2951.

Found, B. and Rogers, D. (1995). Contemporary issues in forensic handwriting examination. a discussion of key issues in the wake of the starzecpyzel decision. *Journal of Forensic Document Examination*, 8:1–31.

Fuentes, M., Garcia-Salicetti, S., and Dorizzi, B. (2002). On line signature verification: Fusion of a hidden markov model and a neural network via a support vector machine. In Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on, pages 253–258. IEEE.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 388 of 404

Galbally, J., Fierrez, J., Martinez-Diaz, M., and Ortega-Garcia, J. (2009a). Evaluation of brute-force attack to dynamic signature verification using synthetic samples. In Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on, pages 131–135. IEEE.

Galbally, J., Fierrez, J., Martinez-Diaz, M., and Ortega-Garcia, J. (2009b). Improving the enrollment in dynamic signature verfication with synthetic samples. In Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on, pages 1295–1299. IEEE.

Galbally, J., Fierrez, J., Martinez-Diaz, M., Ortega-Garcia, J., Plamondon, R., and O'Reilly, C. (2010). Kinematical analysis of synthetic dynamic signatures using the sigma-lognormal model. In Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on, pages 113–118. IEEE.

Galbally, J., Fierrez, J., Ortega-Garcia, J., and Plamondon, R. (2012a). Synthetic on-line signature generation. part ii: Experimental validation. *Pattern Recognition*, 45(7):2622–2632.

Galbally, J., Plamondon, R., Fierrez, J., and Ortega-Garcia, J. (2012b). Synthetic on-line signature generation. part i: Methodology and algorithms. *Pattern Recognition*, 45(7):2610–2621.

Garcia-Salicetti, S., Beumier, C., Chollet, G., Dorizzi, B., Les Jardins, J. L., Lunter, J., Ni, Y., and Petrovska-Delacrétaz, D. (2003). Biomet: A multimodal person authentication database including face, voice, fingerprint, hand and signature modalities. In International Conference on Audio-and Video-based Biometric Person Authentication, pages 845–853. Springer.

Gomez-Barrero, M., Galbally, J., Fierrez, J., Ortega-Garcia, J., and Plamondon, R. (2015). Enhanced on-line signature verification based on skilled forgery detection using sigma-lognormal features. In *Biometrics (ICB), 2015 International Conference on*, pages 501–506. IEEE.

Griechisch, E., Malik, M. I., and Liwicki, M. (2014). Online signature verification based on kolmogorov-smirnov distribution distance. In Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on, pages 738–742. IEEE.

Guest, R. (2006). Age dependency in handwritten dynamic signature verification systems. *Pattern Recognition Letters*, 27(10):1098–1104.

Guest, R. and Fairhurst, M. (2006). Sample selection for optimising signature enrolment. In Tenth International Workshop on Frontiers in Handwriting Recognition. Suvisoft.

Guest, R. M. (2004). The repeatability of signatures. In Frontiers in Handwriting Recognition, 2004. IWFHR-9 2004. Ninth International Workshop on, pages 492–497. IEEE.

Guru, D. and Prakash, H. (2009). Online signature verification and recognition: An approach based on symbolic representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(6):1059–1073.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 389 of 404

Hongo, Y., Muramatsu, D., and Matsumoto, T. (2005). Adaboost-based on-line signature verifier. In *Defense and Security*, pages 373–380. International Society for Optics and Photonics.

Houmani, N., Garcia-Salicetti, S., and Dorizzi, B. (2008). A novel personal entropy measure confronted with online signature verification systems' performance. In Biometrics: Theory, Applications and Systems, 2008. BTAS 2008. 2nd IEEE International Conference on, pages 1–6. IEEE.

Houmani, N., Garcia-Salicetti, S., and Dorizzi, B. (2009). On assessing the robustness of pen coordinates, pen pressure and pen inclination to time variability with personal entropy. In Biometrics: Theory, Applications, and Systems, 2009. BTAS'09. IEEE 3rd International Conference on, pages 1–6. IEEE.

Huang, K. and Yan, H. (1995). On-line signature verification based on dynamic segmentation and global and local matching. *Optical Engineering*, 34(12):3480–3487.

Huang, K. and Yan, H. (2003). Stability and style-variation modeling for on-line signature verification. *Pattern Recognition*, 36(10):2253–2270.

Huber, R. A. and Headrick, A. M. (1999). *Handwriting identification: facts and fundamentals*. CRC press Boca Raton.

Ibrahim, M. T., Kyan, M., Khan, M. A., and Guan, L. (2010). On-line signature verification using 1-d velocity-based directional analysis. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3830–3833. IEEE.

Igarza, J. J., Goirizelaia, I., Espinosa, K., Hernáez, I., Méndez, R., and Sánchez, J. (2003). Online handwritten signature verification using hidden markov models. In Iberoamerican Congress on Pattern Recognition, pages 391–399. Springer.

Igarza, J. J., Gómez, L., Hernáez, I., and Goirizelaia, I. (2004). Searching for an optimal reference system for on-line signature verification based on (x, y) alignment. In *Biometric Authentication*, pages 519–525. Springer.

Impedovo, D. and Pirlo, G. (2008). Automatic signature verification: the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(5):609–635.

Impedovo, D. and Pirlo, G. (2010). On-line signature verification by stroke-dependent representation domains. In Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on, pages 623–627. IEEE.

Impedovo, D., Pirlo, G., Mangini, F. M., Barbuzzi, D., Rollo, A., Balestrucci, A., Impedovo, S., Sarcinella, L., O'Reilly, C., and Plamondon, R. (2013). Writing generation model for health care neuromuscular system investigation. In International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics, pages 137–148. Springer.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 390 of 404

Impedovo, S. and Pirlo, G. (2007). Verification of handwritten signatures: an overview. In Image Analysis and Processing, 2007. ICIAP 2007. 14th International Conference on, pages 191–196. IEEE.

ISO (2007). Information technology biometric data interchange formats part 7: Signature/sign time series data. *ISO/IEC FCD 19794-7*.

Jain, A. K., Griess, F. D., and Connell, S. D. (2002). On-line signature verification. *Pattern recognition*, 35(12):2963–2972.

Kamel, N. S., Sayeed, S., and Ellis, G. A. (2008). Glove-based approach to online signature verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6):1109–1113.

Kashi, R., Hu, J., Nelson, W., and Turin, W. (1998). A hidden markov model approach to online handwritten signature verification. *International Journal on Document Analysis and Recognition*, 1(2):102–109.

Kashi, R. S., Hu, J., Nelson, W., and Turin, W. (1997). On-line handwritten signature verification using hidden markov model features. In Document Analysis and Recognition, 1997., Proceedings of the Fourth International conference on, volume 1, pages 253–257. IEEE.

Khan, M. K., Khan, M. A., Khan, M. A., and Ahmad, I. (2006). On-line signature verification by exploiting inter-feature dependencies. In Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, volume 2, pages 796–799. IEEE.

Kholmatov, A. and Yanikoglu, B. (2005). Identity authentication using improved online signature verification method. *Pattern recognition letters*, 26(15):2400–2408.

KOMIYA, Y., Ohishi, T., and Matsumoto, T. (2001). A pen input on-line signature verifier integrating position, pressure and inclination trajectories. *IEICE transactions on information and systems*, 84(7):833–838.

Leclerc, F. and Plamondon, R. (1994). Automatic signature verification: The state of the art1989–1993. *International Journal of Pattern Recognition and Artificial Intelligence*, 8(03):643–660.

Lee, J., Yoon, H.-S., Soh, J., Chun, B. T., and Chung, Y. K. (2004). Using geometric extrema for segment-to-segment characteristics comparison in online signature verification. *Pattern Recognition*, 37(1):93–103.

Lee, L. L., Berger, T., and Aviczer, E. (1996). Reliable online human signature verification systems. *IEEE Transactions on pattern analysis and machine intelligence*, 18(6):643–647.

Lei, H. and Govindaraju, V. (2005). A comparative study on the consistency of features in on-line signature verification. *Pattern Recognition Letters*, 26(15):2483–2489.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy Page 391 of 404

Lejtman, D. Z. and George, S. E. (2001). On-line handwritten signature verification using wavelets and back-propagation neural networks. In Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on, pages 992–996. IEEE.

Li, B., Wang, K., and Zhang, D. (2004). On-line signature verification based on pca (principal component analysis) and mca (minor component analysis). In *Biometric Authentication*, pages 540–546. Springer.

Longstaff, M. G. and Heath, R. A. (2006). Spiral drawing performance as an indicator of fine motor function in people with multiple sclerosis. *Human movement science*, 25(4):474–491.

Maiorana, E., Campisi, P., Fierrez, J., Ortega-Garcia, J., and Neri, A. (2010). Cancelable templates for sequence-based biometrics with application to on-line signature recognition. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(3):525–538.

Martens, R. and Claesen, L. (1998). Incorporating local consistency information into the online signature verification process. *International Journal on Document Analysis and Recognition*, 1(2):110–115.

Martínez Díaz, M., Fiérrez, J., and Ortega-García, J. (2008). Incorporating signature verification on handheld devices with user-dependent hidden markov models.

Munich, M. E. and Perona, P. (2002). Visual input for pen-based computers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):313–328.

Munich, M. E. and Perona, P. (2003). Visual identification by signature tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):200–217.

Muramatsu, D. and Matsumoto, T. (2003). An hmm on-line signature verification algorithm. In International Conference on Audio-and Video-Based Biometric Person Authentication, pages 233–241. Springer.

Nabeshima, S., Yamamoto, S., Agusa, K., and Taguchi, T. (1995). Memo-pen: a new input device. In Conference companion on Human factors in computing systems, pages 256–257. ACM.

Nakanishi, I., Nishiguchi, N., Itoh, Y., and Fukui, Y. (2004). On-line signature verification based on discrete wavelet domain adaptive signal processing. In *Biometric Authentication*, pages 584–591. Springer.

Nanni, L. (2006). An advanced multi-matcher method for on-line signature verification featuring global features and tokenised random numbers. *Neurocomputing*, 69(16):2402–2406.

Nanni, L., Maiorana, E., Lumini, A., and Campisi, P. (2010). Combining local, regional and global matchers for a template protected on-line signature verification system. *Expert Systems with Applications*, 37(5):3676–3684.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 392 of 404

Nelson, W., Turin, W., and Hastie, T. (1994). Statistical methods for on-line signature verification. *International Journal of Pattern Recognition and Artificial Intelligence*, 8(03):749–770.

Nyssen, E., Sahli, H., and Zhang, K. (2002). A multi-stage online signature verification system. *Pattern Analysis & Applications*, 5(3):288–295.

OReilly, C. and Plamondon, R. (2009). Development of a sigma–lognormal representation for on-line signatures. *Pattern Recognition*, 42(12):3324–3337.

OReilly, C. and Plamondon, R. (2011). Impact of the principal stroke risk factors on human movements. *Human movement science*, 30(4):792–806.

O'Reilly, C. and Plamondon, R. (2012a). Design of a neuromuscular disorders diagnostic system using human movement analysis. In Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on, pages 787–792. IEEE.

O'Reilly, C. and Plamondon, R. (2012b). Looking for the brain stroke signature. In Pattern Recognition (ICPR), 2012 21st International Conference on, pages 1811–1814. IEEE.

Ortega-Garcia, J., Fierrez, J., Alonso-Fernandez, F., Galbally, J., Freire, M. R., Gonzalez-Rodriguez, J., Garcia-Mateo, C., Alba-Castro, J.-L., Gonzalez-Agulla, E., Otero-Muras, E., et al. (2010). The multiscenario multienvironment biosecure multimodal database (bmdb). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):1097–1111.

Ortega-Garcia, J., Fiérrez-Aguilar, J., Martin-Rello, J., and Gonzalez-Rodriguez, J. (2003a). Complete signal modeling and score normalization for function-based dynamic signature verification. In International Conference on Audio-and Video-Based Biometric Person Authentication, pages 658–667. Springer.

Ortega-Garcia, J., Fierrez-Aguilar, J., Simon, D., Gonzalez, J., Faundez-Zanuy, M., Espinosa, V., Satue, A., Hernaez, I., Igarza, J.-J., Vivaracho, C., et al. (2003b). Mcyt baseline corpus: a bimodal biometric database. *IEE Proceedings-Vision, Image and Signal Processing*, 150(6):395–401.

Parizeau, M. and Plamondon, R. (1989). What types of scripts can be used for personal identity verification? *Computer Recognition and Human Production of Handwriting*, pages 77–90.

Parizeau, M. and Plamondon, R. (1990). A comparative analysis of regional correlation, dynamic time warping, and skeletal tree matching for signature verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):710–717.

Phillips, P. J., Martin, A., Wilson, C. L., and Przybocki, M. (2000). An introduction evaluating biometric systems. *Computer*, 33(2):56–63.

Pirlo, G. (1994). Algorithms for signature verification. In *Fundamentals in Handwriting Recognition*, pages 435–454. Springer.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

Pirlo, G., Cuccovillo, V., Diaz-Cabrera, M., Impedovo, D., and Mignone, P. (2015a). Multidomain verification of dynamic signatures using local stability analysis. *IEEE Transactions on Human-Machine Systems*, 45(6):805–810.

Pirlo, G., Diaz, M., Ferrer, M. A., Impedovo, D., Occhionero, F., and Zurlo, U. (2015b). Early diagnosis of neurodegenerative diseases by handwritten signature analysis. In International Conference on Image Analysis and Processing, pages 290–297. Springer.

Plamondon, R. (1994). *Progress in automatic signature verification*, volume 13. World Scientific.

Plamondon, R. (1995). A kinematic theory of rapid human movements: Part iii. kinetic outcomes. *Biological Cybernetics*, 72(4):295–307.

Plamondon, R. and Djioua, M. (2006). A multi-level representation paradigm for handwriting stroke generation. *Human movement science*, 25(4):586–607.

Plamondon, R. and Lorette, G. (1989). Automatic signature verification and writer identificationthe state of the art. *Pattern recognition*, 22(2):107–131.

Plamondon, R., Pirlo, G., and Impedovo, D. (2014). Online signature verification. In *Handbook of Document Image Processing and Recognition*, pages 917–947. Springer.

Plamondon, R. and Srihari, S. N. (2000). Online and off-line handwriting recognition: a comprehensive survey. *IEEE Transactions on pattern analysis and machine intelligence*, 22(1):63–84.

Plimmer, B., Grundy, J., Hosking, J., and Priest, R. (2006). Inking in the ide: Experiences with pen-based design and annotatio. In Visual Languages and Human-Centric Computing, 2006. VL/HCC 2006. IEEE Symposium on, pages 111–115. IEEE.

Qu, T., El Saddik, A., and Adler, A. (2003). Dynamic signature verification system using stroked based features. In Haptic, Audio and Visual Environments and Their Applications, 2003. HAVE 2003. Proceedings. The 2nd IEEE Internatioal Workshop on, pages 83–88. IEEE.

Rabasse, C., Guest, R., and Fairhurst, M. (2007). A method for the synthesis of dynamic biometric signature data. In Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on, volume 1, pages 168–172. IEEE.

Rosenblum, S., Samuel, M., Zlotnik, S., Erikh, I., and Schlesinger, I. (2013). Handwriting as an objective tool for parkinsons disease diagnosis. *Journal of neurology*, 260(9):2357–2361.

Salicetti, S. G., Houmani, N., and Dorizzi, B. (2008). A client-entropy measure for on-line signatures. In Biometrics Symposium, 2008. BSYM'08, pages 83–88. IEEE.

Schmidt, C. and Kraiss, K.-F. (1997). Establishment of personalized templates for automatic signature verification. In Document Analysis and Recognition, 1997., Proceedings of the Fourth International Conference on, volume 1, pages 263–267. IEEE.

Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 394 of 404

Schomaker, L. R. and Plamondon, R. (1990). The relation between pen force and pen-point kinematics in handwriting. *Biological Cybernetics*, 63(4):277–289.

Shafiei, M. M. and Rabiee, H. R. (2003). A new online signature verification algorithm using variable length segmentation and hidden markov models. In Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on, pages 443–446. IEEE.

Swanepoel, J. and Coetzer, J. (2014). Feature weighted support vector machines for writer-independent on-line signature verification. In Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on, pages 434–439. IEEE.

Tolba, A. (1999). Glovesignature: A virtual-reality-based system for dynamic signature verification. *Digital Signal Processing*, 9(4):241–266.

Tolosana, R., Vera-Rodriguez, R., Ortega-Garcia, J., and Fierrez, J. (2015). Preprocessing and feature selection for improved sensor interoperability in online biometric signature verification. *IEEE Access*, 3:478–489.

Uludag, U., Pankanti, S., Prabhakar, S., and Jain, A. K. (2004). Biometric cryptosystems: issues and challenges. Proceedings of the IEEE, 92(6):948–960.

Van, B. L., Garcia-Salicetti, S., and Dorizzi, B. (2007). On using the viterbi path along with hmm likelihood information for online signature verification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 37(5):1237–1247.

Van Gemmert, A., Adler, C. H., and Stelmach, G. (2003). Parkinson's disease patients undershoot target size in handwriting and similar tasks. *Journal of Neurology, Neurosurgery & Psychiatry*, 74(11):1502–1508.

Vielhauer, C. (2005). A behavioural biometrics. *Public Service Review: European Union*, 9:113–115.

Wang, D., Zhang, Y., Yao, C., Wu, J., Jiao, H., and Liu, M. (2010). Toward force-based signature verification: A pen-type sensor and preliminary validation. *IEEE Transactions on Instrumentation and Measurement*, 59(4):752–762.

Wessels, T. and Omlin, C. W. (2000). A hybrid system for signature verification. In Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on, volume 5, pages 509–514. IEEE.

Wibowo, C. P., Thumwarin, P., and Matsuura, T. (2014). On-line signature verification based on forward and backward variances of signature. In Information and Communication Technology, Electronic and Electrical Engineering (JICTEE), 2014 4th Joint International Conference on, pages 1–5. IEEE.

Wirotius, M., Ramel, J.-Y., and Vincent, N. (2005). Comparison of point selection for characterizing on-line signature. In *Defense and Security*, pages 307–313. International Society for Optics and Photonics.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 395 of 404

Woch, A., Plamondon, R., and O'Reilly, C. (2011). Kinematic characteristics of successful movement primitives in young and older subjects: a delta-lognormal comparison. *Hum. Mov. Sci*, 30:1–17.

Wolf, F., Basu, T., Dutta, P. K., Vielhauer, C., Oermann, A., and Yegnanarayana, B. (2006). A cross-cultural evaluation framework for behavioral biometric user authentication. In *From Data and Information Analysis to Knowledge Engineering*, pages 654–661. Springer.

Wu, Q.-Z., Jou, I.-C., and Lee, S.-Y. (1997a). On-line signature verification using lpc cepstrum and neural networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 27(1):148–153.

Wu, Q.-Z., Lee, S.-Y., and Jou, I.-C. (1997b). On-line signature verification based on split-and-merge matching mechanism. *Pattern Recognition Letters*, 18(7):665–673.

Wu, Q.-Z., Lee, S.-Y., Jou, I.-C., et al. (1998). On-line signature verification based on logarithmic spectrum. *Pattern Recognition*, 31(12):1865–1871.

Yan, J. H., Rountree, S., Massman, P., Doody, R. S., and Li, H. (2008). Alzheimer's disease and mild cognitive impairment deteriorate fine movement control. *Journal of Psychiatric Research*, 42(14):1203–1212.

Yang, L., Widjaja, B., and Prasad, R. (1995). Application of hidden markov models for signature verification. *Pattern recognition*, 28(2):161–170.

Yeung, D.-Y., Chang, H., Xiong, Y., George, S., Kashi, R., Matsumoto, T., and Rigoll, G. (2004). Svc2004: First international signature verification competition. In *Biometric Authentication*, pages 16–22. Springer.

Yoon, H., Lee, J., and Yang, H. (2002). An online signature verification system using hidden markov model in polar space. In Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on, pages 329–333. IEEE.

Yue, K. and Wijesoma, W. (2000). Improved segmentation and segment association for on-line signature verification. In Systems, Man, and Cybernetics, 2000 IEEE International Conference on, volume 4, pages 2752–2756. IEEE.

Zou, M., Tong, J., Liu, C., and Lou, Z. (2003). On-line signature verification using local shape analysis. In Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on, pages 314–318. IEEE.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 396 of 404

ABOUT THE EDITORS

Byron Leite Dantas Bezerra
Universidade de Pernambuco (Brazil)
Associate Professor

Byron Leite completed his Ph.D. with an emphasis in Artificial Intelligence in the Federal University of Pernambuco (Brazil) in 2008. He began his activities as Professor of the Polytechnic School at the University of Pernambuco in 2009. He is currently an Associate Professor, Researcher, and Coordinator of the Post-graduation in Computer Engineering at the University of Pernambuco. At the leadership of the Pattern Recognition Research Group, Byron has developed dozens of research and technological innovation projects, bringing together students and researchers from his network of collaborations. He has experience in Computer Science, working mainly on the following topics: document processing, handwriting recognition, gesture recognition, recommendation systems, filtering and information retrieval. Since 2010 it has been developing important collaborations with researchers and excellence groups with international relevance. Through fruitful collaborations with companies, Byron has developed several research and development projects in its research areas, highlighting technologies for processing digital image documents for the private and public market. Through these partnerships with the high-tech companies, Byron has contributed to the design and improvement of document capture and imaging systems, forms processing, handwriting recognition and signature verification software, which, after 10 years of use, have already processed more than 1 billion documents.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

Cleber Zanchettin
Universidade Federal de Pernambuco (Brazil)
Adjunct Professor

Cleber Zanchettin received the Ph.D. degree in computer science from the Universidade Federal de Pernambuco, Recife, Brazil, in 2008. He is currently a Professor and a Technical Reviewer with the Center of Informatics, Universidade Federal de Pernambuco. He has authored over 60 papers in international refereed journals and conferences in pattern recognition, artificial neural networks (ANNs), and intelligent systems. He serves as reviewer for many international journals including IEEE-Transactions on Systems, Man and Cybernetics, IEEE-Transactions on Neural Networks and Learning Systems, Elsevier-Talanta, Elsevier-Applied Soft Computing, Elsevier-Neurocomputing, Elsevier-Engineering Applications of Artificial Intelligence, Springer-Neural Computing & Applications, Springer-International Journal of Machine Learning and Cybernetics, Wiley-International Journal of Adaptive Control and Signal Processing, World Scientific-International Journal of Computational Intelligence and Applications, etc.. He has been in the scientific committee and program committee of many International Conferences like NCTA, IJCCI, IJCNN, ICONIP, IBERAMIA, BRACIS, ENIA, SBC, and SBRN. C. Zanchettin was editor of the Special Issue “Advances in Intelligent Systems — Selected papers from the 2012 Brazilian Symposium on Neural Networks (SBRN 2012)” of the Elsevier-Neurocomputing (2014) and of the Special Issue "Feature and algorithm selection with Hybrid Intelligent Techniques" of the International Journal of Hybrid Intelligent Systems (2012). His current research interests include hybrid neural systems and applications of ANNs.

Alejandro Héctor Toselli
Universitat Politècnica de València (Spain)
Researcher

Alejandro H. Toselli received the M.S. degree in electrical engineering from Universidad Nacional de Tucuman (Argentina) in 1997 and the Ph.D. degree in Computer Science from Universitat Politècnica de València (Spain) in 2004. From 2005 to 2008, he worked as a research member of the "Instituto Tecnológico de Informática" at the Universitat Politècnica de València (Spain), participating in different research projects with technological transfer to industry. From 2008 to 2009, he was a Postdoctoral Fellow at "Institut de Recherche en Informatique et Systèmes Aléatoires" (IRISA, Rennes France), in the "Recognition and interpretation of Images and Documents" Research Group (IMADOC). From 2010 until now, he has a full time position as researcher at the "Pattern Recognition and Human Language Research

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 398 of 404

Center" (PRHLT) at the Universitat Politècnica de València (Spain), being active in different European Research Projects, like EU READ and HIMANIS, focused mainly on the topic of Handwritten Text Recognition. His current research interest lies in the general subject of Document Analysis and Processing under traditional and user-interactive Pattern Recognition view. Particularly, he focus on Handwritten Text Recognition and Keyword Spotting on Historical Documents by employing recognition methodologies such as neural networks and hidden Markov models. He has teaching experience in imparting different tutorial courses about the ``Handwritten Text Recognition" and other related topics.

Giuseppe Pirlo
Università degli Studi di Bari Aldo Moro (Italy)
Associate Professor

Giuseppe Pirlo is Associate Professor at the Department of Computer Science of the University of Bari. He developed and directed several scientific projects and published more than two-hundred-fifty papers on international journals, scientific books and conference proceedings. G. Pirlo is associate editor for IEEE Trans. on Human-Machine Systems. He serves as reviewer for many international journals including IEEE T-PAMI, IEEE T-FS, IEEE T-SMC, IEEE T-EC, IEEE T-IP, IEEE T-IFS, IEEE-SPL, Pattern Recognition, IJDAR, IPL, etc.. He was the general chair of the “International Workshop on Image-based Smart City Applications” (Genoa 2015), “International Workshop on Emerging Aspects in Handwriting Signature Processing” (Naples 2013), and the general co-chair of the “International Conference on Frontiers in Handwriting Recognition” (Bari 2012). He has been in the scientific committee and program committee of many International Conferences in the field of Computer Science, Pattern Recognition and Signal processing, like ICPR, ICDAR, ICFHR, IWFHR, ICIAP, VECIM, and CISMA. G. Pirlo was editor of the Special Issue “Handwriting Recognition and Other PR Applications” of the Pattern Recognition Journal (2014) and of the Special Issue "Handwriting Biometrics" of the IET Biometrics Journal (2014). G. Pirlo was also the co-editor of the book “Advances in Digital Handwritten Signature Processing – A Human Artefact for e-Society”, published by Springer n 2014.

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 399 of 404

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

INDEX

A

academic performance, 333
accuracy, vii, 14, 17, 18, 20, 25, 27, 35, 64, 66, 67, 68, 70, 71, 72, 74, 75, 77, 79, 88, 97, 120, 155, 158, 182, 193, 211, 216, 221, 222, 268, 279, 283, 288, 347, 375
alphabet, 334
ancestors, 196
anisotropy, 332
annotation, 198, 259, 261, 262, 264, 265
assessment, 49, 156, 278, 334, 335, 341, 343, 354, 370
authentication, 345, 346, 347, 354, 359, 379, 381, 386
authenticity, 364, 368, 369
automate, 192
automaticity, 343
automation, 273, 334

B

basic education, 339
Bentham, 66, 73, 75, 76, 82, 83, 84, 97, 103, 104, 105, 106, 107, 108, 109, 115, 116, 121, 122, 124, 125, 126, 133, 152, 310, 311, 312
binarization, 5, 33, 35, 36, 37, 38, 42, 44, 45, 46, 47, 49, 52, 53, 54, 55, 56, 57, 58, 59, 62, 63, 64, 65, 69, 84, 86, 87, 89, 90, 91, 92, 152, 157, 158, 159, 163, 164, 194, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 244, 259, 262, 265, 267, 271, 272, 273, 274, 299, 301, 302, 303, 304, 316
binary decision, 78

C

CDF 9/7 Wavelet Transform, 5, 97, 102, 108
comparative analysis, 364, 383
complexity, 19, 20, 26, 95, 114, 150, 189, 193, 194, 198, 199, 201, 284, 375, 376, 377
comprehension, 27, 34
computation, 30, 60, 63, 96, 120, 125, 145, 162, 175, 192, 196, 201, 220, 303, 314
computer, 3, 11, 24, 30, 45, 81, 95, 113, 114, 144, 169, 273, 278, 364
computer systems, 3, 113
construction, 19, 47, 49, 149, 184, 209, 230, 232, 233, 234, 235, 236, 237, 238, 239, 252, 272, 337
cultural heritage, vii, 227, 277

D

database, 25, 70, 75, 78, 79, 80, 81, 83, 84, 88, 103, 105, 106, 115, 116, 124, 126, 140, 141, 146, 149, 150, 151, 152, 153, 163, 220, 233, 251, 269, 270, 273, 287, 294, 311, 313, 346, 347, 348, 349, 350, 351, 353, 359, 360, 361, 370, 379, 383
datasets, vii, 5, 22, 23, 24, 25, 57, 62, 65, 71, 76, 82, 91, 97, 149, 150, 151, 153, 154, 158, 160, 182, 190, 200, 213, 297, 298, 302, 306, 310, 314, 346, 347, 348, 351, 352, 354, 356, 358
decoding, 82, 96, 99, 118, 133, 139, 141, 148, 280, 282, 283, 284, 287, 289, 290, 291, 292, 293, 306, 336
deformation, 19, 20, 21, 22, 25, 27, 31, 193, 303
degradation, 4, 35, 46, 53, 58, 63, 64, 85, 95, 171, 297
Dense SIFT (DSIFT), 78
detection, 12, 36, 47, 53, 54, 57, 63, 66, 67, 68, 70, 86, 99, 175, 205, 206, 213, 223, 234, 251, 254, 271, 306, 319, 329, 360, 371, 379
dimensionality, 102, 211
Discrete Wavelet Transform (DWT), 100, 371
disorder, 335, 342
displacement, 13, 19, 20, 366, 372
distortions, 19, 21, 27, 58, 286
diversity, 5, 98, 160
dominance, 97, 142, 175, 209
drawing, 46, 79, 86, 322, 382
dyes, 227, 230
dyslexia, 335, 343

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy

E

economic disadvantage, 342
education, 331, 339, 340
educational psychology, 341
elementary school, 242, 341
encoding, 81, 191, 192, 195, 199, 213, 288, 300
equalization, 36, 37, 43, 44, 55
ergonomics, 282, 290
evolution, 6, 27, 135, 136
expert decisions, 371
expert systems, 375
extraction, vii, 5, 33, 39, 47, 54, 64, 70, 71, 75, 85, 88, 89, 91, 92, 96, 97, 98, 100, 105, 106, 122, 125, 126, 142, 153, 182, 201, 211, 212, 213, 224, 230, 246, 247, 249, 250, 254, 263, 267, 270, 271, 272, 273, 274, 275, 286, 300, 302, 304, 305, 307, 347, 356, 358, 364, 371, 372
extractionx method, 71, 246, 272, 275
extracts, 71, 141, 298

F

feature detectors, 145
feature selection, 385
feature(s) extraction, vii, 5, 39, 64, 75, 96, 97, 100, 105, 106, 122, 123, 125, 126, 142, 182, 211, 212, 213, 224, 230, 246, 247, 249, 250, 251, 254, 263, 267, 270, 272, 273, 274, 286, 299, 300, 305, 314, 364, 366, 371
filters, 21, 101, 102, 103, 194
fingerprints, 349
formation, 214, 216, 222, 326, 341
formula, 170, 174, 207, 208, 257
Fourier analysis, 371

G

grades, 333, 342
graph, 11, 61, 67, 76, 77, 90, 120, 125, 174, 175, 181, 196, 197, 198, 199, 208, 209, 280, 300, 303, 319
grouping, 69, 70, 181

H

handheld devices, 365, 372, 382
handwriting, vii, 4, 5, 6, 9, 12, 14, 17, 27, 28, 29, 71, 82, 86, 89, 90, 96, 97, 98, 99, 105, 109, 110, 113, 114, 115, 117, 121, 142, 143, 144, 145, 146, 147, 148, 149, 150, 153, 154, 155, 156, 160, 161, 162, 163, 164, 171, 182, 191, 207, 208, 209, 211, 212, 213, 214, 216, 222, 224, 225, 270, 279, 290, 313, 316, 319, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 333, 334, 339, 340, 341, 342, 343, 347, 349, 361, 363, 365, 376, 377, 378, 384, 385
handwriting analysis, vii
handwriting synthesis, 153, 162
handwritten character recognition, 31, 97, 230, 245
handwritten signature, vii, 345, 346, 347, 348, 349, 353, 354, 358, 361, 363, 364, 365, 375, 377, 380, 381, 382, 384
handwritten signature verification systems, vii
handwritten text recognition (HTR), vii, 57, 73, 95, 97, 99, 101, 103, 105, 107, 109, 110, 111, 144, 147, 277, 294, 295
hardware design for on-line HTR, vii
HCC, 384
high school, 242, 333, 339
histogram, 34, 36, 40, 43, 55, 212, 220, 246, 248, 249, 252, 254, 301, 304, 305
historical archives, vii
historical data, 57, 67, 70, 71, 73, 152
historical handwritten documents, 57, 59, 62, 66, 67, 73, 74, 87, 89, 92, 95, 161, 279, 315
history, 307
HTR systems, vii, 98, 279, 286
HTR workflow, vii
HTR-related applications, vii
HTR-related topics, vii

I

IAM, 71, 97, 115, 121, 122, 124, 125, 126, 131, 132, 133, 139, 140, 141, 142, 143, 146, 151, 310, 312, 313
image analysis, 228, 265, 274, 297
image thresholding, 35, 52, 53, 55, 164
individual character, 78, 286
individual differences, 341
individuals, 16, 211, 345, 348, 349, 350, 353, 370, 375
international competition, 199, 311
interoperability, 347, 354, 358, 365, 376, 377, 385
issues, 4, 15, 16, 25, 28, 33, 41, 48, 68, 171, 190, 191, 227, 333, 365, 378, 385

K

keyword spotting, vii, 6, 57, 79, 80, 81, 82, 83, 84, 85, 87, 90, 91, 93, 152, 297, 314, 315, 316, 317, 318, 319

L

languages, 12, 79, 163, 169, 184, 212, 214, 246, 251, 305, 313
latency, 322, 326
latent semantic indexing (LSI), 251, 252, 263, 268, 269, 306

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 402 of 404

lattices, 121, 135, 141, 291
LEAF, 5, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 242, 245, 246, 251, 258, 259, 260, 261, 262, 263, 264, 265, 266, 268, 270
learning, 6, 9, 11, 12, 15, 17, 21, 22, 27, 30, 31, 61, 64, 67, 79, 80, 81, 82, 84, 92, 93, 107, 135, 145, 155, 161, 169, 200, 206, 242, 251, 297, 299, 309, 314, 317, 319, 333, 339, 340, 341, 346, 368
learning disabilities, 341
learning process, 107, 242
learning task, 17
literacy, 341
local thresholds, 63

M

machine learning, 9, 12, 29, 30, 32, 54, 71, 85, 86, 114, 143, 150, 153, 155, 223
Markov chain, 82
mathematical formulae recognition, vii
mathematics, 172, 206, 208
Maurdor, 146, 152
memory, 13, 14, 156, 182, 200, 332, 339
MEMS, 321, 332
mental retardation, 335
mobile device, 6, 329, 347, 358, 365
modern society, 345, 363
modifications, 15, 16, 19, 24, 158, 376
motor control, 364, 376
MSW, 39, 40, 41, 42
multidimensional, 17, 75, 162
multimodal handwriting speech recognition, vii
multiresolution analysis (MRA), 100

N

narratives, 341
neodymium, 328
neurodegenerative diseases, 384
neuropsychology, 341
neuroscience, 12, 364
NFI, 351
NIST, 23, 24, 25, 26, 150, 151, 308, 317

O

operations, 6, 7, 12, 13, 35, 36, 43, 44, 61, 101, 125, 152, 157, 158, 193, 212, 285
optimization, 10, 31, 53, 70, 71, 89, 110, 143, 145

P

pattern recognition, 30, 53, 126, 144, 149, 169, 190, 223, 273, 346
pen pressure, 4, 350, 380
personal identity, 364
physical characteristics, 242, 244, 363
physiological factors, 337
physiology, 30, 286
preservation, 48, 235, 238
PRImA, 153
primary school, 333, 341, 342
prior knowledge, 17, 18, 28, 67, 174
probability, 54, 98, 118, 120, 125, 135, 147, 148, 176, 178, 179, 180, 181, 183, 184, 185, 186, 187, 188, 201, 202, 203, 279, 280, 282, 285, 303, 370
probability distribution, 303
professionals, 333, 334, 335, 336
programming, 25, 187, 188, 280, 287, 305
project, 29, 104, 111, 147, 161, 162, 228, 242, 269, 270, 351
propagation, 21, 27, 30, 31, 148, 372, 382
psychology, 341, 342
public schools, 334
PUMA, 307, 308, 309

Q

QoS, 375
query, 6, 7, 71, 79, 80, 81, 83, 251, 252, 254, 255, 256, 257, 258, 261, 263, 265, 268, 297, 298, 299, 300, 303, 304, 305, 306, 307, 308, 314, 318, 346
query-by-example (QBE), 79
query-by-string (QBS), 80
query-by-String (QbS), 84

R

reading, 3, 5, 6, 15, 28, 34, 66, 159, 175, 279, 336
realism, 288
reality, 28, 228, 365, 385
receiver operating characteristic (ROC), 370
receptive field, 22, 23
recognition systems, vii, 5, 6, 80, 82, 98, 122, 164, 176, 209, 286, 287, 289, 290
reconstruction, 54, 60, 100, 101
recurrence, 117, 130, 132, 137, 142
recurrent neural network, 13, 82, 84, 113, 114, 117, 121, 142, 145, 182, 205, 206
reference system, 380
reinforcement learning, 24
rejection, 176, 298, 370
religion, 73, 104, 228
researchers, vii, viii, 4, 27, 28, 68, 83, 150, 222, 228, 245, 297, 346, 354, 364
resolution, 23, 44, 45, 58, 85, 100, 102, 103, 151, 158, 181, 189, 236, 237, 321, 322, 329, 348, 349, 350, 352, 354
resources, 28, 100, 192, 337
RIMES, 143, 152

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 403 of 404

rules, 28, 66, 73, 175, 176, 184, 185, 189, 201, 204, 212, 216, 218, 300, 301, 371

S

school, 333, 334, 335
school performance, 333
science, 3, 144, 271, 315, 340, 342, 364, 382, 383, 384
scripting language recognition, vii
scripts, 5, 77, 83, 169, 172, 211, 212, 222, 224, 228, 245, 246, 251, 273, 383
short-term memory, 30, 73, 87, 145, 162, 213
SIGMA, 350, 358, 359
skimming, 378
society, 3, 271, 273, 315, 363, 374
spatial information, 62, 176, 179, 180, 188, 300, 303
speech, vii, 6, 12, 31, 96, 97, 111, 114, 121, 125, 135, 143, 144, 145, 146, 147, 148, 212, 225, 278, 279, 280, 281, 282, 283, 284, 286, 290, 291, 293, 294, 295, 301, 305, 334, 335
speech processing, 301
spelling, 73, 304, 335, 337, 342
statistic test, 336
statistical, 30, 31, 32, 109, 110, 154, 177, 208, 211, 224, 225, 294, 383
statistics, 54, 150, 289, 360, 371
stroke, 4, 34, 35, 36, 39, 40, 43, 44, 45, 48, 54, 62, 63, 64, 81, 89, 96, 156, 170, 171, 177, 180, 181, 182, 189, 196, 197, 198, 202, 203, 204, 205, 210, 212, 213, 214, 215, 216, 217, 218, 219, 220, 222, 223, 234, 239, 289, 303, 347, 354, 356, 371, 376, 377, 380, 383, 384
stroke width, 34, 39, 40, 43, 44, 45, 48
structural contrast, 39, 40, 41, 42, 53
structure, 6, 11, 13, 15, 21, 24, 70, 79, 97, 98, 103, 104, 108, 122, 150, 153, 157, 164, 172, 173, 174, 175, 176, 177, 178, 179, 184, 187, 191, 192, 197, 199, 201, 205, 207, 252, 315, 321, 322, 329, 331, 364
structuring, 44, 237
style(s), 4, 16, 22, 66, 73, 76, 77, 95, 98, 104, 170, 211, 214, 246, 251, 297, 304, 313, 319, 375, 380
symbols and connections, 153

T

techniques, vii, 4, 5, 6, 9, 12, 16, 17, 23, 26, 27, 28, 53, 55, 57, 58, 59, 60, 61, 62, 63, 64, 65, 71, 73, 74, 81, 88, 95, 96, 98, 99, 105, 150, 153, 159, 160, 164, 224, 270, 271, 273, 278, 298, 305, 342, 363, 364, 365, 368, 371, 372, 373, 374, 375, 378
technological advances, 4
technologies, 5, 57, 95, 270
technology, vii, 6, 95, 153, 277, 279, 322, 331, 332, 345, 360, 365, 377, 381
test data, 80, 82, 288, 289
testing, 11, 22, 23, 27, 75, 77, 78, 82, 84, 97, 105, 179, 220, 222, 259, 263, 300, 342, 352
text image preprocessing, vii
texture, 5, 35, 48, 62, 67, 86, 109, 234, 314, 346
thresholding, vii, 5, 33, 34, 35, 36, 38, 39, 40, 42, 44, 47, 48, 52, 53, 54, 55, 62, 63, 64, 69, 86, 102, 122, 158, 164, 236, 274, 301, 303
time constraints, 27
time periods, 13, 376
time series, 381
time warp, 90, 317, 372, 383
traits, 345, 349, 363, 378
transactions, 29, 55, 305, 345, 360, 377, 379, 381
transcription, 5, 49, 61, 73, 78, 80, 87, 95, 98, 104, 107, 113, 133, 141, 142, 143, 148, 151, 152, 162, 229, 234, 236, 237, 259, 277, 278, 279, 280, 281, 282, 283, 289, 290, 291, 293, 294, 314, 319, 341
transcripts, 107
transformation, 10, 17, 19, 100, 103, 106, 108, 210, 371, 372

V

validation, 11, 21, 24, 26, 75, 77, 78, 105, 107, 115, 116, 122, 124, 125, 140, 141, 200, 201, 203, 217, 218, 259, 267, 269, 282, 340, 379, 385
valuation, 143, 146, 239, 240, 241, 244
variables, 45, 119, 179, 289, 336, 337, 338, 376
variations, 15, 22, 24, 26, 27, 39, 42, 45, 46, 63, 124, 216, 230, 257, 287, 297, 298, 304, 354
visual acuity, 43
visual perception, 42, 53, 55
visual system, 44
vocabulary, 74, 75, 77, 90, 96, 105, 109, 114, 122, 124, 125, 126, 135, 139, 140, 141, 142, 146, 147, 155, 277, 279, 287, 289

W

wavelet, 60, 92, 97, 98, 100, 101, 102, 104, 106, 111, 112, 382
wavelet analysis, 112
wavelet transform, 5, 97, 98, 100, 101, 102, 108, 109, 367, 371
word recognition, 16, 75, 77, 88, 90, 96, 110, 144, 152, 212, 223, 318
writing process, 4, 6, 342, 346, 363, 364
writing tasks, 334

Cerence Operating Company, Ex. 2011
Samsung Electronics Co., Ltd., etc. v. Cerence Operating Company, IPR2024-01267
Complimentary Contributor Copy
Page 404 of 404