# DECLARATION OF JUNE ANN MUNFORD

1. My name is June Ann Munford. I am over the age of 18, have personal knowledge of the facts set forth herein, and am competent to testify to the same.

2. I earned a Master of Library and Information Science (MLIS) from the University of Wisconsin-Milwaukee in 2009. I have over ten years of experience in the library/information science field. Beginning in 2004, I have served in various positions in the public library sector including Assistant Librarian, Youth Services Librarian and Library Director. I have attached my Curriculum Vitae as Appendix CV.

3. During my career in the library profession, I have been responsible for materials acquisition for multiple libraries. In that position, I have cataloged, purchased and processed incoming library works. That includes purchasing materials directly from vendors, recording publishing data from the material in question, creating detailed material records for library catalogs and physically preparing that material for circulation. In addition to my experience in acquisitions, I was also responsible for analyzing large collections of library materials, tailoring library records for optimal catalog

search performance and creating lending agreements between libraries during my time as a Library Director.

4. I am fully familiar with the catalog record creation process in the library sector. In preparing a material for public availability, a library catalog record describing that material would be created. These records are typically written in Machine Readable Catalog (herein referred to as “MARC”) code and contain information such as a physical description of the material, metadata from the material’s publisher, and date of library acquisition. In particular, the 008 field of the MARC record is reserved for denoting the date of creation of the library record itself. As this typically occurs during the process of preparing materials for public access, it is my experience that an item’s MARC record indicates the date of an item’s public availability.

5. Typically, in creating a MARC record, a librarian would gather various bits of metadata such as book title, publisher and subject headings among others and assign each value to a relevant numerical field. For example, a book’s physical description is tracked in field 300 while title/attribution is tracked in field 245. The 008 field of the MARC record is reserved for denoting the creation of the library record itself. As this is the only date reflecting the inclusion of said materials within the library’s collection, it is my experience

that an item's 008 field accurately indicates the date of an item's public availability.

6. I have reviewed Exhibit 1013, "A Statistical Model for General Contextual Object Recognition" by Peter Carbonetto, et al. as published in *Proceedings of the 8th European Conference on Computer Vision, 2004*.

7. Attached hereto as Appendix CARBONETTO01 is a true and correct copy of the MARC record for *Proceedings of the 8th European Conference on Computer Vision, 2004* as held by the Case Western University library. I secured this record myself from the library's public catalog. The MARC record contained within Appendix CARBONETTO01 accurately describes the title, author, publisher, and ISBN number of *Proceedings of the 8th European Conference on Computer Vision, 2004*. In comparing Exhibit 1013 to Appendix CARBONETTO01, it is my determination that Exhibit 1013 is a true and correct copy of "A Statistical Model for General Contextual Object Recognition" as published in *Proceedings of the 8th European Conference on Computer Vision, 2004* by Peter Carbonetto, et al..

8. The 008 field of the MARC record in Appendix CARBONETTO01 indicates the date of record creation. The 008 field of Appendix

CARBONETTO01 indicates the Case Western University library first acquired these proceedings as of May 24, 2004. Considering this information, it is my determination that *Proceedings of the 8th European Conference on Computer Vision, 2004* and therefore "A Statistical Model for General Contextual Object Recognition" was made available to the public shortly after its initial acquisition as of May 24, 2004.

9. I have reviewed Exhibit 1014, "Hello! My name is...Buffy - Automatic Naming of Characters in TV Video" by M. Everingham, et al. as published in *Proceedings of the British Machine Vision Conference 2006*.

10. Attached hereto as Appendix EVERINGHAM01 is a true and correct copy of the MARC record for *Proceedings of the British Machine Vision Conference 2006* as held by the Library of Trinity College Dublin. I secured this record myself from the library's public catalog. The MARC record contained within Appendix EVERINGHAM01 accurately describes the title, author, publisher, and ISBN number of *Proceedings of the British Machine Vision Conference 2006*. In comparing Exhibit 1014 to Appendix EVERINGHAM01, it is my determination that Exhibit 1014 is a true and correct copy of "Hello! My name is...Buffy - Automatic Naming of

Characters in TV Video" as published in *Proceedings of the British Machine Vision Conference 2006* by M. Everingham, et al.

11. The 008 field of the MARC record in Appendix EVERINGHAM01 indicates the date of record creation. The 008 field of Appendix EVERINGHAM01 indicates the Library of Trinity College Dublin first acquired these proceedings as of January 5, 2007. Considering this information, it is my determination that *Proceedings of the British Machine Vision Conference 2006* and therefore "Hello! My name is...Buffy - Automatic Naming of Characters in TV Video" was made available to the public by the Library of Trinity College Dublin shortly after its initial acquisition as of January 5, 2007.

12. I am familiar with the Internet Archive, a digital library formally certified by the State of California as a public library. Among other services that the Internet Archive makes available to the general public is the Wayback Machine, an online archive. The Internet Archive's Wayback Machine service archives webpages as of a certain capture date to track changes in the web over time. The Internet Archive has been in operation as a nonprofit library since 1996 and has hosted the Wayback Machine service since its inception in 2001. During my time as a librarian, I frequently used the

Internet Archive's Wayback Machine for research and instruction purposes. This includes teaching instructional classes on using the Wayback Machine to library patrons and using the Wayback Machine to research reference inquiries that require hard-to-find online resources. I consider the Internet Archive's recordskeeping to be as rigorous and detailed as other formal library recordskeeping practices such as MARC records, OCLC records and Dublin Core.

13. Attached hereto as Appendix EVERINGHAM02 is a screen capture of the Internet Archive Wayback Machine entry for http://www.macs.hw.ac.uk/bmvc2006/. I secured these screen captures myself from https://web.archive.org/web/20061010155114/http://www.macs.hw.ac.uk/bmvc2006/. This record was captured by the Internet Archive as of October 10, 2006.

14. Attached hereto as Appendix EVERINGHAM02 is a capture of a web page entitled "BMVC 2006" as hosted on Heriot Watt University. This page describes the 2006 British Machine Vision Conference and contains a link advertising the conference proceedings for download. The presence of this link indicates the organizers of the 2006 British Machine Vision Conference

intended to make these proceedings publicly available via this webpage as of October 10, 2006.

15. I have reviewed Exhibit 1015, "Recognizing Products: A per-exemplar multi-label image classification approach" by M. George, et al. as published in *Proceedings of the 2014 Computer Vision European Conference*.

16. Attached hereto as Appendix GEORGE01 is a true and correct copy of the MARC record for *Proceedings of the 2014 Computer Vision European Conference* as held by the Penn State University library. I secured this record myself from the library's public catalog. The MARC record contained within Appendix GEORGE01 accurately describes the title, author, publisher, and ISBN number of *Proceedings of the 2014 Computer Vision European Conference.*

17. Attached hereto as Appendix GEORGE02 is a true and correct copy of "Recognizing Products: A per-exemplar multi-label image classification approach" as published in *Proceedings of the 2014 Computer Vision European Conference* as held by the Penn State University library. I secured this file myself from the library's collection. In comparing Exhibit 1015 to Appendix GEORGE02, it is my determination that Exhibit 1015 is a true and

correct copy of “Recognizing Products: A per-exemplar multi-label image classification approach” as published in *Proceedings of the 2014 Computer Vision European Conference* by M. George, et al.

18. The 008 field of the MARC record in Appendix GEORGE01 indicates the date of record creation. The 008 field of Appendix GEORGE01 indicates the Penn State University library first acquired these proceedings as of August 13, 2014. Considering this information, it is my determination that *Proceedings of the 2014 Computer Vision European Conference* and therefore “Recognizing Products: A per-exemplar multi-label image classification approach” was made available to the public shortly after its initial acquisition as of August 13, 2014.

19. I have reviewed Exhibit 1016, “Product counting using images with application to robot-based retail stock assessment” by Nishant Kejriwal, et al. as published in *Proceedings of the 2015 IEEE International Conference on Technologies for Practical Robot Applications*.

20. Attached hereto as Appendix KEJRIWAL01 is a true and correct copy of the MARC record for *Proceedings of the 2015 IEEE International Conference on Technologies for Practical Robot Applications* as held by the University

of Washington library. I secured this record myself from the library's public catalog. The MARC record contained within Appendix KEJRIWAL01 accurately describes the title, author, publisher, and ISBN number of *Proceedings of the 2015 IEEE International Conference on Technologies for Practical Robot Applications*. In comparing Exhibit 1016 to Appendix KEJRIWAL01, it is my determination that Exhibit 1016 is a true and correct copy of "Product counting using images with application to robot-based retail stock assessment" as published in *Proceedings of the 2015 IEEE International Conference on Technologies for Practical Robot Applications* by Nishant Kejriwal, et al..

21. The 008 field of the MARC record in Appendix KEJRIWAL01 indicates the date of record creation. The 008 field of Appendix KEJRIWAL01 indicates the University of Washington library first acquired these proceedings as of September 29, 2015. Considering this information, it is my determination that *Proceedings of the 2015 IEEE International Conference on Technologies for Practical Robot Applications* and therefore "Product counting using images with application to robot-based retail stock assessment" was made available to the public shortly after its initial acquisition as of September 29, 2015.

22. I have reviewed Exhibit 1017, “Remote Retail Monitoring and Stock Assessment using Mobile Robots” by Swagat Kumar, et al. as published in *Proceedings of the 2014 IEEE Conference on Technologies for Practical Robot Applications*.

23. Attached hereto as Appendix KUMAR01 is a true and correct copy of the MARC record for *Proceedings of the 2014 IEEE Conference on Technologies for Practical Robot Applications* as held by the University of Washington library. I secured this record myself from the library’s public catalog. The MARC record contained within Appendix KUMAR01 accurately describes the title, author, publisher, and ISBN number of *Proceedings of the 2014 IEEE Conference on Technologies for Practical Robot Applications*. In comparing Exhibit 1017 to Appendix KUMAR01, it is my determination that Exhibit 1017 is a true and correct copy of “Remote Retail Monitoring and Stock Assessment using Mobile Robots” as published in *Proceedings of the 2014 IEEE Conference on Technologies for Practical Robot Applications* by Swagat Kumar, et al.

24. The 008 field of the MARC record in Appendix KUMAR01 indicates the date of record creation. The 008 field of Appendix KUMAR01 indicates the University of Washington library first acquired these proceedings as of

October 10, 2014. Considering this information, it is my determination that *Proceedings of the 2014 IEEE Conference on Technologies for Practical Robot Applications* and therefore “Remote Retail Monitoring and Stock Assessment using Mobile Robots” was made available to the public shortly after its initial acquisition as of October 10, 2014.

25. I have reviewed Exhibit 1018, "Recognizing Groceries in situ Using in vitro training data" by Michele Merler, et al. as published in *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition*.

26. Attached hereto as Appendix MERLER01 is a true and correct copy of the MARC record for *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition* as held by the Penn State University library. I secured this record myself from the library’s public catalog. The MARC record contained within Appendix MERLER01 accurately describes the title, author, publisher, and ISBN number of *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition*.

27. Attached hereto as Appendix MERLER02 is a true and correct copy of "Recognizing Groceries in situ Using in vitro training data" as published in *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern*

*Recognition* as held by the Penn State University library. I secured this file myself from the library's collection. In comparing Exhibit 1018 to Appendix MERLER02, it is my determination that Exhibit 1018 is a true and correct copy of "Recognizing Groceries in situ Using in vitro training data" as published in *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition* by Michele Merler, et al.

28. The 008 field of the MARC record in Appendix MERLER01 indicates the date of record creation. The 008 field of Appendix MERLER01 indicates the Penn State University library first acquired these proceedings as of February 26, 2007. Considering this information, it is my determination that *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition* and therefore "Recognizing Groceries in situ Using in vitro training data" was made available to the public shortly after its initial acquisition as of March 26, 2007.

29. I have reviewed Exhibit 1022 *Machine Learning* by Tom M. Mitchell.

30. Attached hereto as Appendix MITCHELL01 is a true and correct copy of the MARC record for *Machine Learning* as held by the University of Pittsburgh library. I secured this record myself from the library's public catalog. The

MARC record contained within Appendix MITCHELL01 accurately describes the title, author, publisher, and ISBN number of *Machine Learning.*

31. The 008 field of the MARC record in Appendix MITCHELL01 indicates the date of record creation. The 008 field of Appendix MITCHELL01 indicates University of Pittsburgh library first acquired this book as of February 11, 1997. Considering this information, it is my determination that *Machine Learning* was made available to the public shortly after its initial acquisition as of February 11, 1997.

32. I have reviewed Exhibit 1019 “Ontological Supervision for Fine Grained Classification of Street View Storefronts” by Yair Movshovitz-Attias, et al. as published in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*.

33. Attached hereto as Appendix MOVSHOVITZ01 is a true and correct copy of the MARC record for *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition* as held by the Penn State University library. I secured this record myself from the library’s public catalog. The MARC record contained within Appendix MOVSHOVITZ01

accurately describes the title, author, publisher, and ISBN number of *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition.*

34. Attached hereto as Appendix MOVSHOVITZ02 is a true and correct copy of "Ontological Supervision for Fine Grained Classification of Street View Storefronts" as published in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition* as held by the Penn State University library. I secured this file myself from the library's collection. In comparing Exhibit 1019 to Appendix MOVSHOVITZ02, it is my determination that Exhibit 1019 is a true and correct copy of "Ontological Supervision for Fine Grained Classification of Street View Storefronts" as published in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition* by Yair Movshovitz-Attias, et al.

35. The 008 field of the MARC record in Appendix MOVSHOVITZ01 indicates the date of record creation. The 008 field of Appendix MOVSHOVITZ01 indicates the Penn State University library first acquired these proceedings as of October 14, 2015. Considering this information, it is my determination that *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition* and therefore "Ontological Supervision for Fine

Grained Classification of Street View Storefronts" was made available to the public shortly after its initial acquisition as of October 14, 2015.

36. I have reviewed Exhibit 1020 "Toward retail product recognition on grocery shelves" by Gul Varol, et al. as published in *Proceedings of the 2014 International Conference on Graphic and Image Processing*.

37. Attached hereto as Appendix VAROL01 is a true and correct copy of the MARC record for *Proceedings of the 2014 International Conference on Graphic and Image Processing* as held by the University of Washington library. I secured this record myself from the library's public catalog. The MARC record contained within Appendix VAROL01 accurately describes the title, author, publisher, and ISBN number of *Proceedings of the 2014 International Conference on Graphic and Image Processing*. In comparing Exhibit 1020 to Appendix VAROL01, it is my determination that Exhibit 1020 is a true and correct copy of "Toward retail product recognition on grocery shelves" as published in *Proceedings of the 2014 International Conference on Graphic and Image Processing* by Gul Varol, et al..

38. The 008 field of the MARC record in Appendix VAROL01 indicates the date of record creation. The 008 field of Appendix VAROL01 indicates the

University of Washington library first acquired these proceedings as of April 8, 2015. Considering this information, it is my determination that *Proceedings of the 2014 International Conference on Graphic and Image Processing* and therefore "Toward retail product recognition on grocery shelves" was made available to the public shortly after its initial acquisition as of April 8, 2015.

39. I have reviewed Exhibit 1021 "Searching the Web with Mobile Image for Location Recognition" by Tom Yeh, et al. as published in *Proceedings of the 2004 Computer Vision and Pattern Recognition Conference*.

40. Attached hereto as Appendix YEH01 is a true and correct copy of the MARC record for *Proceedings of the 2004 Computer Vision and Pattern Recognition Conference* as held by the Penn State University library. I secured this record myself from the library's public catalog. The MARC record contained within Appendix YEH01 accurately describes the title, author, publisher, and ISBN number of *Proceedings of the 2004 Computer Vision and Pattern Recognition Conference.*

41. Attached hereto as Appendix YEH02 is a true and correct copy of "Searching the Web with Mobile Image for Location Recognition" as

published in *Proceedings of the 2004 Computer Vision and Pattern Recognition Conference* as held by the Penn State University library. I secured this file myself from the library's collection. In comparing Exhibit 1021 to Appendix YEH02, it is my determination that Exhibit 1021 is a true and correct copy of "Searching the Web with Mobile Image for Location Recognition" as published in *Proceedings of the 2004 Computer Vision and Pattern Recognition Conference* by Tom Yeh, et al.

42. The 008 field of the MARC record in Appendix YEH01 indicates the date of record creation. The 008 field of Appendix YEH01 indicates the Penn State University library first acquired these proceedings as of September 3, 2004. Considering this information, it is my determination that *Proceedings of the 2004 Computer Vision and Pattern Recognition Conference* and therefore "Searching the Web with Mobile Image for Location Recognition" was made available to the public shortly after its initial acquisition as of September 3, 2004.

43. I have been retained on behalf of the Petitioner to provide assistance in the above-illustrated matter in establishing the authenticity and public availability of the documents discussed in this declaration. I am being compensated for my services in this matter at the rate of $200.00 per hour

plus reasonable expenses. My statements are objective, and my compensation does not depend on the outcome of this matter.

44. I declare under penalty of perjury that the foregoing is true and correct. I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code.

Dated: 6/18/2024



June Ann Munford

**Appendix CV**

June A. Munford
Curriculum Vitae

**Education**

University of Wisconsin-Milwaukee - MS, Library & Information Science, 2009
Milwaukee, WI

- Coursework included cataloging, metadata, data analysis, library systems, management strategies and collection development.
- Specialized in library advocacy, cataloging and public administration.

Grand Valley State University - BA, English Language & Literature, 2008
Allendale, MI

- Coursework included linguistics, documentation and literary analysis.
- Minor in political science with a focus in local-level economics and government.

**Professional Experience**

Researcher / Expert Witness, October 2017 – present
Freelance ● Pittsburgh, Pennsylvania & Grand Rapids, Michigan

- Material authentication and public accessibility determination. Declarations of authenticity and/or public accessibility provided upon research completion. Experienced with appeals and deposition process.

- Research provided on topics of public library operations, material publication history, digital database services and legacy web resources.

- Past clients include Alston & Bird, Arnold & Porter, Baker Botts, Fish & Richardson, Erise IP, Irell & Manella, O'Melveny & Myers, Perkins-Coie, Pillsbury Winthrop Shaw Pittman and Slayden Grubert Beard.

Library Director, February 2013 - March 2015
Dowagiac District Library ● Dowagiac, Michigan

- Executive administrator of the Dowagiac District Library. Located in

Southwest Michigan, this library has a service area of 13,000, an annual operating budget of over $400,000 and total assets of approximately $1,300,000.

● Developed careful budgeting guidelines to produce a 15% surplus during the 2013-2014 & 2014-2015 fiscal years while being audited.

● Using this budget surplus, oversaw significant library investments including the purchase of property for a future building site, demolition of existing buildings and building renovation projects on the current facility.

● Led the organization and digitization of the library's archival records.

● Served as the public representative for the library, developing business relationships with local school, museum and tribal government entities.

● Developed an objective-based analysis system for measuring library services - including a full collection analysis of the library's 50,000+ circulating items and their records.

November 2010 - January 2013
Librarian & Branch Manager, Anchorage Public Library ● Anchorage, Alaska

● Headed the 2013 Anchorage Reads community reading campaign including event planning, staging public performances and creating marketing materials for mass distribution.

● Co-led the social media department of the library's marketing team, drafting social media guidelines, creating original content and instituting long-term planning via content calendars.

● Developed business relationships with The Boys & Girls Club, Anchorage School District and the US Army to establish summer reading programs for children.

June 2004 - September 2005, September 2006 - October 2013
Library Assistant, Hart Area Public Library
Hart, MI

● Responsible for verifying imported MARC records and original MARC

cataloging for the local-level collection as well as the Michigan Electronic Library.

● Handled OCLC Worldcat interlibrary loan requests & fulfillment via ongoing communication with lending libraries.

**Professional Involvement**

Alaska Library Association - Anchorage Chapter
- Treasurer, 2012

Library Of Michigan
- Level VII Certification, 2008
- Level II Certification, 2013

Michigan Library Association Annual Conference 2014
- New Directors Conference Panel Member

Southwest Michigan Library Cooperative
- Represented the Dowagiac District Library, 2013-2015

**Professional Development**

Library Of Michigan Beginning Workshop, May 2008
Petoskey, MI
- Received training in cataloging, local history, collection management, children's literacy and reference service.

Public Library Association Intensive Library Management Training, October 2011
Nashville, TN
- Attended a five-day workshop focused on strategic planning, staff management, statistical analysis, collections and cataloging theory.

Alaska Library Association Annual Conference 2012 - Fairbanks, February 2012
Fairbanks, AK
- Attended seminars on EBSCO advanced search methods, budgeting, cataloging, database usage and marketing.

**Depositions**

2019 ● Fish & Richardson

Apple v. Qualcomm (IPR2018-001281, 39521-00421IP, IPR2018-01282 and 39521-00421IP2)

2019 ● Erise IP

Implicit, LLC v. Netscout Systems, Inc (Civil Action No. 2:18-cv-53-JRG)

2019 ● Perkins-Coie

Adobe Inc. v. RAH Color Technologies LLC (Cases IPR2019-00627, IPR2019-00628, IPR2019-00629 and IPR2019-00646)

2020 ● O'Melveny & Myers

Maxell, Ltd. v. Apple Inc. (Case No. 5:19-cv-00036-RWS)

2021 ● Pillsbury Winthrop Shaw Pittman LLP

Intel v. SRC (IPR2020-1449)

2022 ● Perkins-Coie

Realtek v. Future Link (IPR2021-01182)

2023 ● Fish & Richardson

Neuroderm Ltd. v. Abbvie, Inc (Case No. PGR2022-00040)

2023 ● Fish & Richardson

Nearmap US Inc. v. Pictometry International Corp. (IPR2022-00735)

2023 ● Fish & Richardson

Samsung Electronics v. MemoryWeb LLC (Case No. 39843-0136PS1)

2023 ● Pillsbury Winthrop Shaw Whitman LLP

Gravel Rating Systems v. Costco Wholesale Corp. (Civil Action No. 4:21-cv-149-ALM)

**Limited Case and Clientele History**

Alston & Bird

- Ericsson

  v. Collision Communications (IPR2022-01233)

- Nokia

  v. Neptune Subsea, Xtera (Case No. 1:17-cv-01876)

- Universal Electronics Inc

  v. Roku Inc (IPR2022-00818)

Arnold & Porter

- Ivantis

  v. Glaukos (Case No. 8:18-cv-00620)

- Samsung

  v. Jawbone (Case No. 2:21-cv-00186)

Benesch Friedlander Coplan & Aronoff

- Voyis

  v. Cathx (Case No.  5:21-cv-00077-RWS)

Erise I.P.

- Apple
  - v. Ericsson Inc. (IPR2022-00715)
  - v. Future Link Systems (IPRs 6317804, 6622108, 6807505, and 7917680)
  - v. INVT (Case No. 20-1881)
  - v. Navblazer LLC (IPR2020-01253)
  - v. Qualcomm (IPR2018-001281, 39521-00421IP, IPR2018-01282, 39521-00421IP2)
  - v. Quest Nettech Corp (Case No. 2:19-cv-00118-JRG)
  - v. Telefonaktiebolaget LM Ericsson (IPR2022-00275)
- Fanduel
  - v. CGT (Case No. 19-1393)
- Garmin
  - v. Phillips North America LLC (Case No. 2:19-cv-6301-AB-KS)
- Netscout
  - v. Longhorn HD LLC (Case No. 2:20-cv-00349)
  - v. Implicit, LLC (Civil Action No. 2:18-cv-53-JRG)
- Sony Interactive Entertainment LLC
  - v. Bot M8 LLC (IPR2020-01288)
  - v. Infernal Technology LLC (Case No. 2:19-CV-00248-JRG)
- Unified Patents

v. GE Video Compression (Civil Action No. 2:19-cv-248)

Fish & Richardson

- Apple
  - v. AliveCor (3:21-cv-03958)
  - v. LBS Innovations (Case No. 2:19-cv-00119-JRG-RSP)
  - v. Koss Corporation (IPR2021-00305)
  - v. Masimo (IPR 50095-0012IP1, 50095-0012IP2, 50095-0013IP1, 50095-0013IP2, 50095-0006IP1, 50095-0135IP1)
  - v. Neonode (Case No. 21-cv-08872-EMC)
  - v. Qualcomm (IPR2018-001281, 39521-00421IP, IPR2018-01282, 39521-00421IP2)
- Dell
  - v. Neo Wireless (IPR2022-00616)
- Dish Network
  - v. Entropic Communications, LLC (Case No. 2:2023-CV-01043)
  - v. Realtime Adaptive Streaming (Case No. 1:17-CV-02097-RBJ)
  - v. TQ Delta LLC (Case No. 18-1798)
- Evapco Dry Cooling
  - v. SPG Dry Cooling (IPR2021-00688)
- Genetec
  - v. Sensormatic Electronics (Case No. 1:20-CV-00760)
- Huawei

v. Bell Northern Research LLC (IPR2019-01174)

- Kianxis

  v. Blue Yonder (Case No. 3:20-cv-03636)

- LG Electronics

  v. Bell Northern Research LLC (Case No. 3:18-cv-2864-CAB-BLM)

- Metaswitch

  v. Sonus Networks (IPR2018-01719)

- Microsoft

  v. Throughputer Inc (IPR2022-00757)

- Mom Enterprises

  v. Ddrops Company (Case No. 1:22-cv-00332)

- MLC Intellectual Property

  v. MicronTech (Case No. 3:14-cv-03657-SI)

- Nearmap Inc

  v. EagleView Technologies (IPR2022-01009)

- Neuroderm Ltd.

  v. Abbvie, Inc (Case No. PGR2022-00040)

- Realtek Semiconductor

v. Future Link (IPR2021-01182)

- Quectel
  - v. Koninklijke Philips (Case No. 1:20-cv-01710)

- Samsung
  - v. Aire Technology (IPR2022-00877)
  - v. Bell Northern Research (Case No. 2:19-cv-00286-JRG)
  - v. Communication Technologies Inc (IPR2022-01221)
  - v. Jawbone Innovations (IPR2022-00865)
  - v. MemoryWeb LLC (IPR2022-00885)
  - v. Dodots Licensing Solutions (IPR2023-00701)
  - v. Telefonaktiebolaget LM Ericsson (IPR2021-00615)

- Texas Instruments
  - v. Vantage Micro LLC (IPR2020-01261)

- Xilinx
  - v. Sentient Sensors LCC (Case No. 1:22-cv-00173)

Irell & Manella

- Curium

O'Melveny & Myers

- Apple
  - v. Maxell (Case No. 5:19-cv-00036-RWS)

- Micron Technology Inc.

  v. Besang Inc (IPR2023-00900)

- Samsung

  v. Daedalus Prime LLC (1335 Investigation)

Perkins-Coie

- Heru Industries

  v. The UAB Research Foundation (IPR2022-01148)

- Intel Corporation

  v. BitMICRO (Case No. 5:23-cv-00625)

- Realtek Semiconductor

  v. Future Link (IPR2021-01182)

- Twitter Inc

  v. VOIP-Pal.com (Case No. 3:20-cv-02397-JD)

- TCL Industries

  v. Koninklijke Philips NV (IPR2021-00495, IPR2021-00496 and IPR2021-00497)

Pillsbury Winthrop Shaw Pittman

- Intel

  v. FG SRC LLC (Case No. 6:20-cv-00315)

- Gravel Rating Systems

v. Costco (Case No. 4:21-cv-149)
v. Lowe's Home Centers (Case No. 4:21-cv-150)
v. T-Mobile USA (Case No. 4:21-cv-152)
v. Kohl's Inc. (Case No. 4:21-cv-258)
v. Under Armor (Case No. 4:21-cv-356)

Quinn Emanuel

- Wolfspeed

v. The Trustees of Purdue University (Case No. 1:2021-cv-00840)

Stern, Kessler, Goldstein & Fox

- Preliminary Research

Sheppard, Mullin, Richter & Hampton LLP

- Cadence Design Systems Inc.

v. Semiconductor Design Technologies (Case No. 3:23-cv-01001)

**Appendix CARBONETTO01**

```
LEADER 00000cam  2200409Ki 4500
001    55477007
003    OCoLC
005    20041006143251.0
008    040524s2004    gw a     b    001 0 eng d
020    3540219846
020    3540219838
020    354021982X
020    3540219811
040    OSU|erda|cOSU|dOCLCQ|dCWR|beng
049    CWRR
090    QA76|b.L43 no. 3021-3024
111 2  European Conference on Computer Vision|n(8th :|d2004 :
       |cPrague, Czech Republic)
245 10 Computer vision-ECCV 2004 :|b8th European Conference on
       Computer Vision, Prague, Czech Republic, May 11-14, 2004 :
       proceedings /|cTomáš Pajdla, Jiří Matas (editions.)
246 30 ECCV 2004
264  1 Berlin ;|aNew York :|bSpringer,|c[2004]
264  4 |c©2004
300    4 volumes :|billustrations ;|c24 cm
336    text|btxt|2rdacontent
337    unmediated|bn|2rdamedia
338    volume|bnc|2rdacarrier
440  0 Lecture notes in computer science,|x0302-9743 ;|v3021-3024
504    Includes bibliographical references and index
530    Also available via the World Wide Web
650  0 Computer vision|vCongresses
700 1  Pajdla, Tomáš
700 1  Matas, Jiří
856 41 |uhttp://www.springerlink.com/openurl.asp?genre=issue&
       issn=0302-9743&volume=3021|zRestricted to Springer LINK
       subscribers
856 41 |uhttp://www.springerlink.com/openurl.asp?genre=issue&
       issn=0302-9743&volume=3022|zRestricted to Springer LINK
       subscribers
856 41 |uhttp://www.springerlink.com/openurl.asp?genre=issue&
       issn=0302-9743&volume=3023|zRestricted to Springer LINK
       subscribers
856 41 |uhttp://www.springerlink.com/openurl.asp?genre=issue&
       issn=0302-9743&volume=3024|zRestricted to Springer LINK
       subscribers
```

| LOCATION | CALL NO. | STATUS |
|---|---|---|
| KSL Stacks 3rd Floor | QA76.L43 no. 3021-3024 pt. 1 | AVAILABLE |
| KSL Stacks 3rd Floor | QA76.L43 no. 3021-3024 pt. 2 | AVAILABLE |
| KSL Stacks 3rd Floor | QA76.L43 no. 3021-3024 pt. 3 | AVAILABLE |
| KSL Stacks 3rd Floor | QA76.L43 no. 3021-3024 pt. 4 | AVAILABLE |

CWRU

© 2023 Case Western Reserve University

KELVIN SMITH LIBRARY

Campus Location:

11055 Euclid Avenue
Cleveland, OH 44106-7151

*10900 Euclid Ave. Cleveland, Ohio 44106*

216.368.2000
Legal Notice | Privacy Policy

**Phone:** 216.368.3506

**Email:** Ask A Librarian

Site Feedback

CASE WESTERN RESERVE UNIVERSITY

# Library Catalog

Search CWRU...

Borrowing | Research Guides | Case Libraries | Other Sources | Ask A Librarian

Start Over | Add to Bag | MARC Display | Another Search

RECORD # b2897644 Search

Limit search to available items

Search OhioLink



| | |
|---|---|
| Title | **Computer vision-ECCV 2004 : 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004 : proceedings / Tomáš Pajdla, Jiří Matas (editions.)** |
| Conference | **European Conference on Computer Vision (8th : 2004 : Prague, Czech Republic)** |
| Publisher | Berlin ; New York : Springer, [2004] |
| Copyright | ©2004 |

**Click on the following to:**

**Restricted to Springer LINK subscribers**
**Restricted to Springer LINK subscribers**
**Restricted to Springer LINK subscribers**
**Restricted to Springer LINK subscribers**

| LOCATION | CALL NO. | STATUS |
|---|---|---|
| KSL Stacks 3rd Floor Map it | **QA76.L43 no. 3021-3024** pt. 1 | AVAILABLE |
| KSL Stacks 3rd Floor Map it | **QA76.L43 no. 3021-3024** pt. 2 | AVAILABLE |
| KSL Stacks 3rd Floor Map it | **QA76.L43 no. 3021-3024** pt. 3 | AVAILABLE |
| KSL Stacks 3rd Floor Map it | **QA76.L43 no. 3021-3024** pt. 4 | AVAILABLE |

| | |
|---|---|
| Description | 4 volumes : illustrations ; 24 cm |
| Content | text |
| Media | unmediated |
| Format | volume |
| Series | **Lecture notes in computer science,** 0302-9743 ; **3021-3024** |
| Note | Includes bibliographical references and index |
| Additional Format | Also available via the World Wide Web |
| Added Personal Name | **Pajdla, Tomáš** |
| | **Matas, Jiří** |
| Alternate Title | ECCV 2004 |
| Subject | **Computer vision -- Congresses** |
| ISBN | 3540219846<br>3540219838<br>354021982X<br>3540219811 |
| OCLC # | 55477007 |

Google Preview:

*Persistent link to this title*






CWRU

© 2023 Case Western Reserve University

*10900 Euclid Ave. Cleveland, Ohio 44106*

216.368.2000
Legal Notice | Privacy Policy

KELVIN SMITH LIBRARY

**Campus Location:**

11055 Euclid Avenue
Cleveland, OH 44106-7151

**Phone:** 216.368.3506

**Email:** Ask A Librarian

Site Feedback

CWRU LINKS

Apply to CWRU
Give Now
Visit Campus
University Directory

**Appendix EVERINGHAM01**

```
LEADER 00000nam  2200000 a 4500
003    IeDuTC
005    20080714171311.0
008    070105s2006    stka     b    101 0 eng c
015    GBA705836|2bnb
020    1904410146 (set)
040    Uk|beng|cUk
042    ukscp
111 2  British Machine Vision Conference (|d2006 :|cEdinburgh,
       Scotland)
245 10 British Machine Vision Conference 2006 /|cedited by M.J.
       Chantler, E. Trucco & R.B. Fisher.
246 13 BMVC 2006
260    [Edinburgh] :|bBMVA,|cc2006.
300    3 v. :|bill. ;|c24 cm.
500    Cover title.
504    Includes bibliographical references and author's index.
650  0 Computer vision|vCongresses.
653 1  Machine vision
653 1  BMVC
653 2  BMVA
700 1  Chantler, Mike.
700 1  Fisher, R. B.
700 1  Trucco, Emanuele.
710 2  British Machine Vision Association.
935    IeDuTCb14570183
```

| Location | Shelfmark | Status |
|---|---|---|
| Santry Stacks (place request) | PL-467-559 Vol.1 | IN |
| Santry Stacks (place request) | PL-467-560 Vol.2 | IN |
| Santry Stacks (place request) | PL-467-561 Vol.3 | IN |

The Library of Trinity College Dublin
Classic Catalogue





**Conference** British Machine Vision Conference ( 2006 : Edinburgh, Scotland)

Bookmark link for this record

**Title** British Machine Vision Conference 2006 / edited by M.J. Chantler, E. Trucco & R.B. Fisher.

**Publisher** [Edinburgh] : BMVA, c2006.

**Description** 3 v. : ill. ; 24 cm.

| Location | Shelfmark | Status |
|---|---|---|
| Santry Stacks (place request) | PL-467-559 Vol.1 | IN |
| Santry Stacks (place request) | PL-467-560 Vol.2 | IN |
| Santry Stacks (place request) | PL-467-561 Vol.3 | IN |

## More Information

**Note** Cover title.

**Bibliography** Includes bibliographical references and author's index.

**Subject** Computer vision -- Congresses.

**Added Author** Chantler, Mike.
Fisher, R. B.
Trucco, Emanuele.
British Machine Vision Association.

**Other Title** BMVC 2006

**ISBN** 1904410146 (set)

**Appendix EVERINGHAM02**

https://web.archive.org/web/20061010155114/http://www.macs.hw.ac.uk/bmvc2006/ June 12, 2024
INTERNET ARCHIVE WayBack Machine http://www.macs.hw.ac.uk/bmvc2006/ Go SEP OCT DEC
77 captures 6 Feb 2006 - 19 Apr 2022 2005 10 2006 2010 About this capture






# BMVC 2006

Organised by the British Machine Vision Association, the 17th BMVC was held 4-7 September 2006 in Edinburgh.

**Latest news**

*Sep27 Online proceedings available here*

**Speakers:**

Maria Petrou (Imperial College, London)
*Human vision system and data processing - tutorial talk on Monday 4th September - abstract*

David J. Kriegman (University of California, San Diego)
*Vision in the Small: Reconstructing the Structure of Protein Macromolecules from Cryo-Electron Micrographs - Invited talk on Tuesday 5th September - abstract*

Brian Curless (University of Washington)
*Capturing Visual Experiences - Invited talk on Wednesday 6th September - abstract*

**Chairs:** Mike Chantler, Bob Fisher, Manuel Trucco

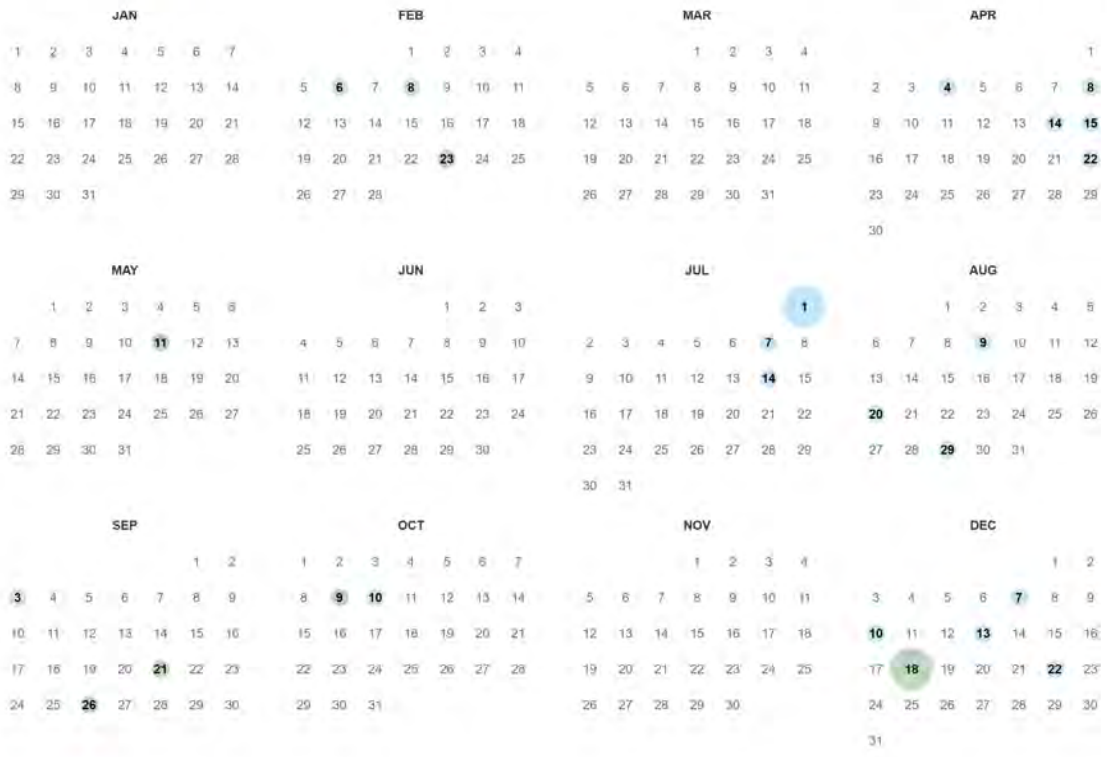
INTERNET ARCHIVE WEB BOOKS VIDEO AUDIO SOFTWARE IMAGES SIGN UP | LOG IN UPLOAD Search
ABOUT BLOG PROJECTS HELP DONATE CONTACT JOBS VOLUNTEER PEOPLE

DONATE INTERNET ARCHIVE WayBackMachine Explore more than 866 billion web pages saved over time
http://www.macs.hw.ac.uk/bmvc2006/

Calendar · Collections · Changes · Summary · Site Map · URLs

Saved **77 times** between February 6, 2006 and April 19, 2022.




## Note

This calendar view maps the number of times **http://www.macs.hw.ac.uk/bmvc2006/** was crawled by the Wayback Machine, *not* how many times the site was actually updated. More info in the FAQ.
Green indicates redirects (3xx).

FAQ | Contact Us | Terms of Service (Dec 31, 2014)

The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & archive-it.org

Your use of the Wayback Machine is subject to the Internet Archive's Terms of Use.

**Appendix GEORGE01**

# MARC View

| | | | | |
|---|---|---|---|---|
| LEADER 02698nam a22005535i 4500 | | | | |
| 001 13209586 | | | | |
| 003 SIRSI | | | | |
| 005 20151213095523.0 | | | | |
| 007 cr nn 008mamaa | | | | |
| 008 140813s2014 gw \| s \|\|\|\| 0\|eng d | | | | |
| 019 | | | a\| MARS | |
| 020 | | | a\| 9783319105901 9\| 978-3-319-10590-1 | |
| 024 | 7 | | a\| 10.1007/978-3-319-10590-1 2\| doi | |
| 040 | | | d\| UtOrBLW | |
| 050 | | 4 | a\| TA1637-1638 | |
| 072 | | 7 | a\| UYT 2\| bicssc | |
| 072 | | 7 | a\| UYQV 2\| bicssc | |
| 072 | | 7 | a\| COM012000 2\| bisacsh | |
| 072 | | 7 | a\| COM016000 2\| bisacsh | |
| 082 | 0 | 4 | a\| 006.6 2\| 23 | |
| 082 | 0 | 4 | a\| 006.37 2\| 23 | |
| 100 | 1 | | a\| Fleet, David. | |
| 245 | 1 | 0 | a\| Computer Vision – ECCV 2014 h\| [electronic resource] : b\| 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I / c\| edited by David Fleet, Tomas Pajdla, Bernt Schiele, Tinne Tuytelaars. | |
| 264 | | 1 | a\| Cham : b\| Springer International Publishing : b\| Imprint: Springer, c\| 2014. | |
| 300 | | | a\| XXVIII, 853 pages 361 illustrations : b\| online resource. | |
| 336 | | | a\| text b\| txt 2\| rdacontent | |
| 337 | | | a\| computer b\| c 2\| rdamedia | |
| 338 | | | a\| online resource b\| cr 2\| rdacarrier | |
| 347 | | | a\| text file b\| PDF 2\| rda | |
| 490 | 1 | | a\| Lecture Notes in Computer Science, x\| 0302-9743 ; v\| 8689 | |
| 520 | | | a\| The seven-volume set comprising LNCS volumes 8689-8695 constitutes the refereed proceedings of the 13th European Conference on Computer Vision, ECCV 2014, held in Zurich, Switzerland, in September 2014. The 363 revised papers presented were carefully reviewed and selected from 1444 submissions. The papers are organized in topical sections on tracking and activity recognition; recognition; learning and inference; structure from motion and feature matching; computational photography and low-level vision; vision; segmentation and saliency; context and 3D scenes; motion and 3D scene analysis; and poster sessions. | |
| 650 | | 0 | a\| Computer science. | |
| 650 | | 0 | a\| Artificial intelligence. | |
| 650 | | 0 | a\| Computer graphics. | |
| 650 | | 0 | a\| Computer vision. | |
| 650 | | 0 | a\| Optical pattern recognition. | |
| 700 | 1 | | a\| Pajdla, Tomáš. | |
| 700 | 1 | | a\| Schiele, Bernt. | |
| 700 | 1 | | a\| Tuytelaars, Tinne. | |
| 710 | 2 | | a\| SpringerLink (Online service) | |
| 773 | 0 | | t\| Springer eBooks | |
| 776 | 0 | 8 | i\| Printed edition: z\| 9783319105895 | |
| 830 | | 0 | a\| Lecture notes in computer science ; v\| 8689. | |
| 856 | 4 | 0 | u\| https://ezaccess.libraries.psu.edu/login?url=http://dx.doi.org/10.1007/978-3-319-10590-1 | |
| 912 | | | a\| ZDB-2-SCS | |
| 912 | | | a\| ZDB-2-LNC | |
| 950 | | | a\| Computer Science (Springer-11645) | |
| 590 | | | a\| AVAILABLE ONLINE TO AUTHORIZED PSU USERS. | |
| 949 | | | a\| Electronic resource w\| ASIS c\| 1 i\| 13209586-1001 l\| ONLINE m\| ONLINE r\| Y s\| Y t\| ONLINE u\| 9/26/2014 o\| .STAFF. Springer2014 imprints since May 20140926 | |

ASK A LIBRARIAN

PennState Copyright @2019 The Pennsylvania State University. All rights

PENN STATE UNIVERSITY LIBRARIES

CONNECT WITH PENN STATE

reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

Legal Statements | PSU Hotlines



(814) 865-6368




# Computer Vision – ECCV 2014 [electronic resource] : 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I / edited by David Fleet, Tomas Pajdla, Bernt Schiele, Tinne Tuytelaars

ASK A LIBRARIAN

Author: Fleet, David

Published: Cham : Springer International Publishing : Imprint: Springer, 2014.

Physical Description: XXVIII, 853 pages 361 illustrations : online resource

Additional Creators: Pajdla, Tomáš, Schiele, Bernt, Tuytelaars, Tinne, and SpringerLink (Online service)

Access Online

ezaccess.libraries.psu.edu

Search inside at Google Books

Series:

Lecture notes in computer science ; 8689

Summary:

The seven-volume set comprising LNCS volumes 8689-8695 constitutes the refereed proceedings of the 13th European Conference on Computer Vision, ECCV 2014, held in Zurich, Switzerland, in September 2014. The 363 revised papers presented were carefully reviewed and selected from 1444 submissions. The papers are organized in topical sections on tracking and activity recognition; recognition; learning and inference; structure from motion and feature matching; computational photography and low-level vision; vision; segmentation and saliency; context and 3D scenes; motion and 3D scene analysis; and poster sessions.

Subject(s):

Computer science
Artificial intelligence
Computer graphics
Computer vision
Optical pattern recognition

ISBN:

9783319105901

Digital File Characteristics:

text file PDF

Note:

AVAILABLE ONLINE TO AUTHORIZED PSU USERS.

Part Of:

Springer eBooks

View MARC record | catkey: 13209586

PennState

Copyright @2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and

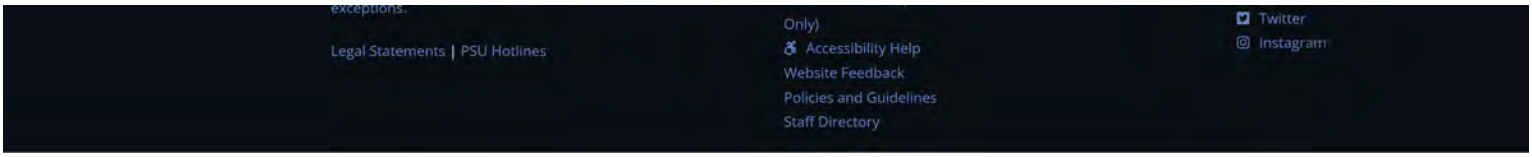
PENN STATE UNIVERSITY LIBRARIES

Libraries Home (814) 865-6368

Libraries Intranet (Staff

CONNECT WITH PENN STATE UNIVERSITY LIBRARIES

Facebook


exceptions.
Legal Statements | PSU Hotlines
Only)
Accessibility Help
Website Feedback
Policies and Guidelines
Staff Directory
Twitter
Instagram

SPRINGER LINK

Log in

Menu Search Cart



# Recognizing Products: A Per-exemplar Multi-label Image Classification Approach

| Conference paper

| pp 440–455 | Cite this conference paper

Download book PDF



Computer Vision – ECCV 2014

(ECCV 2014)

Marian George & Christian Floerkemeier

**Part of the book series:** Lecture Notes in Computer Science ((LNIP,volume 8690))

**Included in the following conference series:**
European Conference on Computer Vision

19k Accesses 57 Citations 3 Altmetric

## Abstract

Large-scale instance-level image retrieval aims at retrieving specific instances of objects or scenes. Simultaneously retrieving multiple objects in a test image adds to the difficulty of the problem, especially if the objects are visually similar. This paper presents an efficient approach for per-exemplar multi-label image classification, which targets the recognition and localization of products in retail store images. We achieve runtime efficiency through the use of discriminative random forests, deformable dense pixel matching and genetic algorithm optimization. Cross-dataset recognition is performed, where our training images are taken in ideal conditions with only one single training image per product label, while the evaluation set is taken using a mobile phone in real-life scenarios in completely different conditions. In addition, we provide a large novel dataset and labeling tools for products image search, to motivate further research efforts on multi-label retail products image classification. The proposed approach achieves promising results in terms of both accuracy and runtime efficiency on 680 annotated images of our dataset, and 885 test images of GroZi-120 dataset. We make our dataset of 8350 different product images and the 680 test images from retail stores with complete annotations available to the wider community.

Download to read the full chapter text

## Similar content being viewed by others




**Automatic Check-Out via Prototype-Based Classifier Learning from Single-Product...**

Chapter | © 2022




**Discrete semi-supervised learning for multi-label image classification and large...**

Article | 19 February 2019



**Multi-instance Learning for Semantic Image Analysis**

Chapter | © 2022

## Keywords

Test Image Training Image Average Precision

Genetic Algorithm Optimization Fisher Vector

*These keywords were added by machine and not by the authors. This process is experimental and the keywords may be updated as the learning algorithm improves.*

## References

1. Bangpeng, Y., Aditya, K., Fei-Fei, L.: Combining randomization and discrimination for fine-grained image categorization. In: CVPR (2011)

Google Scholar

2. Boutella, M.R., Luob, J., Shena, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recognition 37(9), 1755–1771 (2004)

Google Scholar

3. Branson, S., Wah, C., Schroff, F., Babenko, B., Welinder, P., Perona, P., Belongie, S.: Visual recognition with humans in the loop. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 438–451. Springer, Heidelberg (2010)

Chapter Google Scholar

4. Csurka, G., Dance, C., Bray, C., Fan, L.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision (2004)

Google Scholar

5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)

Google Scholar

6. Deep, K., Singh, K.P., Kansal, M.L., Mohan, C.: A real coded genetic algorithm for solving integer and mixed integer optimization problems. Applied Mathematics and Computation 212(2), 505–518 (2009)

Article MATH MathSciNet Google Scholar

7. Duan, G., Huang, C., Ai, H., Lao, S.: Boosting associated pairing comparison features for pedestrian detection. In: ICCV Workshop on Visual Surveillance (2009)

Google Scholar

8. Everingham, M., Gool, L.V., Williams, C., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV 88(2) (2010)

Google Scholar

9. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In: Workshop on Generative-Model Based Vision, CVPR (2004)

Google Scholar

10. Fei-Fei, L., Fergus, R., Torralba, A.: Recognizing and learning object categories. In: ICCV Tutorial (2005)

Google Scholar

11. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE T. Pattern Anal. 32(9), 1627–1645 (2010)

Article Google Scholar

12. Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley (1989)

Google Scholar

13. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Technical report, Caltech (2007)

Google Scholar

14. Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 304–317. Springer, Heidelberg (2008)

Chapter Google Scholar

15. Jing, Y., Baluja, S.: Pagerank for product image search. In: WWW (2008)

Google Scholar

16. Jurie, Y.S.F.: Improving image classifcation using semantic attributes. IJCV 100(1), 59–77 (2012)

Article Google Scholar

17. Kang, F., Jin, R., Sukthankar, R.: Correlated label propagation with application to multi-label learning. In: CVPR (2006)

Google Scholar

18. Khosla, A., Jayadevaprakash, N., Yao, B., Fei-Fei, L.: Novel dataset for fine-grained image categorization. In: First Workshop on Fine-Grained Visual Categorization, CVPR (2011)

Google Scholar

19. Kim, J., Liu, C., Sha, F., Grauman, K.: Deformable spatial pyramid matching for fast dense correspondences. In: CVPR (2013)

Google Scholar

20. Lin, X., Gokturk, B., Sumengen, B., Vu, D.: Visual search engine for product images. In: Multimedia Content Access: Algorithms and Systems II (2008)

Google Scholar

21. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60(2), 91–110 (2004)

Article Google Scholar

22. Merler, M., Galleguillos, C., Belongie, S.: Recognizing groceries in situ using in vitro training data. In: CVPR (2007)

Google Scholar

23. Nilsback, M.E., Zisserman, A.: A visual vocabulary for ower classification. In: CVPR (2006)

Google Scholar

24. Pandey, M., Lazebnik, S.: Scene recognition and weakly supervised object localization with deformable part-based models. In: ICCV (2011)

Google Scholar

25. Perronnin, F., Liu, Y., Sanchez, J., Poirier, H.: Large-scale image retrieval with compressed fisher vectors. In: CVPR (2010)

Google Scholar

26. Sharma, G., Jurie, F., Schmid, C.: Discriminative spatial saliency for image classification. In: CVPR (2012)

Google Scholar

27. Shen, X., Lin, Z., Brandt, J., Wu, Y.: Mobile product image search by automatic query object extraction. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 114–127. Springer, Heidelberg (2012)

Chapter Google Scholar

28. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I, LNCS, vol. 3951, pp. 1–15. Springer, Heidelberg (2006)

Google Scholar

29. Torralba, A., Efros, A.: Unbiased look at dataset bias. In: CVPR (2011)

Google Scholar

30. Tsai, S.S., Chen, D.M., Chandrasekhar, V., Takacs, G., Cheung, N.M., Vedantham, R., Grzeszczuk, R., Girod, B.: Mobile product recognition. In: ACM Multimedia (ACM MM) (2010)

Google Scholar

31. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: CVPR (2010)

Google Scholar

32. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-ucsd birds 200. Technical report cns-tr-201, Caltech (2010)

Google Scholar

33. Winlock, T., Christiansen, E., Belongie, S.: Toward real-time grocery detection for the visually impaired. In: CVAVI (2010)

Google Scholar

34. Zha, Z., Hua, X., Mei, T., Wang, J., Qi, G., Wang, Z.: Joint multi-label multi-instance learning for image classification. In: CVPR (2008)

Google Scholar

35. Zhang, M., Pena, J., Robles, V.: Feature selection for multi-label naive bayes classification. Information Sciences 179(19), 3218–3229 (2009)

Article MATH Google Scholar

# Author information

## Authors and Affiliations

**Department of Computer Science, ETH Zurich, Switzerland**
Marian George & Christian Floerkemeier

# Editor information

## Editors and Affiliations

**Department of Computer Science, University of Toronto, 6 King's College Road, M5H 3S5, Toronto, ON, Canada**
David Fleet

**Faculty of Electrical Engineering, Department of Cybernetics, Czech Technical University in Prague, Technicka 2, 166 27, Prague 6, Czech Republic**
Tomas Pajdla

**Max-Planck-Institut für Informatik, Campus E1 4, 66123, Saarbrücken, Germany**
Bernt Schiele

**KU Leuven, ESAT - PSI, iMinds, Kasteelpark Arenberg, 10, Bus 2441, 3001, Leuven, Belgium**
Tinne Tuytelaars

# Rights and permissions

Reprints and permissions

# Copyright information

© 2014 Springer International Publishing Switzerland

# About this paper

## Cite this paper

George, M., Floerkemeier, C. (2014). Recognizing Products: A Per-exemplar Multi-label Image Classification Approach. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8690. Springer, Cham. https://doi.org/10.1007/978-3-319-10605-2_29

.RIS .ENW .BIB

DOI
https://doi.org/10.1007/978-3-319-10605-2_29

Publisher Name
Springer, Cham

Print ISBN
978-3-319-10604-5

Online ISBN
978-3-319-10605-2

eBook Packages
Computer Science
Computer Science (R0)

## Share this paper

Anyone you share the following link with will be able to read this content:

Get shareable link

Provided by the Springer Nature SharedIt content-sharing initiative

# Publish with us

Policies and ethics

**Appendix GEORGE02**

# Recognizing Products: A Per-exemplar Multi-label Image Classification Approach

Marian George and Christian Floerkemeier

Department of Computer Science
ETH Zurich, Switzerland

**Abstract.** Large-scale instance-level image retrieval aims at retrieving specific instances of objects or scenes. Simultaneously retrieving multiple objects in a test image adds to the difficulty of the problem, especially if the objects are visually similar. This paper presents an efficient approach for per-exemplar multi-label image classification, which targets the recognition and localization of products in retail store images. We achieve runtime efficiency through the use of discriminative random forests, deformable dense pixel matching and genetic algorithm optimization. Cross-dataset recognition is performed, where our training images are taken in ideal conditions with only one single training image per product label, while the evaluation set is taken using a mobile phone in real-life scenarios in completely different conditions. In addition, we provide a large novel dataset and labeling tools for products image search, to motivate further research efforts on multi-label retail products image classification. The proposed approach achieves promising results in terms of both accuracy and runtime efficiency on 680 annotated images of our dataset, and 885 test images of GroZi-120 dataset. We make our dataset of 8350 different product images and the 680 test images from retail stores with complete annotations available to the wider community.

## 1 Introduction

Many image classification techniques try to recognize the object classes present in the image through training multiple binary classifiers or a multi-class classifier using a large number of training images. Improving the performance of such algorithms requires learning the model using as many images as possible that are drawn from the same distribution as the test images [29]. However, in many real-world applications, we face the challenge that the testing images are taken in completely different settings than the training images. For example, in the domain of assisting the visually impaired, or vision for mobile robots, images are very likely to suffer from blur, specularities, unusual viewing angles, a lot of background clutter and very different lighting conditions. Gathering and labeling images that try to mimic the natural environments for which the system is used, is a tedious and very time-consuming task.

Recently, image recognition in the retail products domain has become an interesting research topic due to the remarkable advancements in the capabilities

D. Fleet et al. (Eds.): ECCV 2014, Part II, LNCS 8690, pp. 440–455, 2014.
© Springer International Publishing Switzerland 2014




**Fig. 1.** System overview: (a) Given a test image, (b) we first filter the categories that the test image may belong to, (c) then we match the test image against all images in the filtered categories. (d) An energy function is then optimized given the top-ranked matches to obtain the final list, along with inferred locations, of recognized products.

of mobile phones and mobile vision systems [27,30]. A mobile vision algorithm to recognize products in an image has a wide range of potential applications, ranging from identifying individual products to provide review and price information, to the assisted navigation in supermarkets. Furthermore, mobile retail products recognition can assist the visually impaired in shopping, encouraging them to independently perform daily activities, which promotes their social wellness.

Building a system that parses an image featuring several retail products taken with a smart phone introduces several challenges. This includes the cross-dataset recognition challenges mentioned earlier. Product images available through online shopping websites are taken in ideal studio-like conditions, which are very different from real life images taken with mobile phones in shops, as illustrated in figure 2. These challenges are aggravated when having only one image per product for training and thousands of products (labels) to match against. Due to the increasing number of new products every day, the system also needs to be scalable with no or minimal retraining whenever a new product is introduced. To be applicable in the visually impaired domain, the designed scheme cannot rely on any feedback from the user in improving the retrieved results. The system has to work in a completely autonomous manner. Recognizing grocery products, in particular, is challenging as there are multiple products that have very similar visual appearance except for minor features like the color of the package, or some text describing the product. Finally, runtime efficiency is crucial for mobile vision systems, which makes semantic segmentation or sliding window detection approaches computationally expensive for our problem.

To this end, we designed an efficient per-exemplar multi-label image classification technique (as shown in figure 1) which targets simultaneous recognition and localization of all the individual products in a retail store image. Our algorithm

works in a hierarchical manner, where we first filter the possible labels that a test image may contain through ranking the output of a fine-grained classification model. Then, we perform fast dense pixel matching on the images in our filtered list, and rank the individual products by their matching score. Multi-label image classification is then achieved through minimizing an energy function that correlates the matching score, context and recognition localization results through genetic algorithm global optimization. Our proposed approach is evaluated in a per-exemplar cross-dataset settings, where we show promising results in terms of both accuracy and speed. Our main contributions include:

- A large-scale novel retail products dataset, which contains 8350 different products in 80 different fine-grained categories, and 680 test images taken with a smart phone in natural environments with complete annotations and labeling tools.
- A fast technique to simultaneously recognize and localize all retail products in a test image, which scales to thousands of different labels.
- Automatically infer the total number and approximate locations of objects present in a test image in a single optimization round, unlike other more expensive object detection techniques.
- Experimental evaluation of our proposed system with an analysis of its different components, showing promising results.

## 2 Related Work

General-purpose object detection and recognition techniques [10,11,24] have been extensively used in image classification and retrieval problems. Some approaches [4] use binary classifiers trained on bag-of-words features [4] or compressed Fisher Vectors [25] to obtain a binary decision whether an object class is present or absent in a given test image. Hamming embedding[14] provides binary signatures that refine the matching based on visual words. It has been very successful in instance-level image retrieval domain. Others [11,7] try to detect and localize object classes in the image through extracting dense features like HOG [5] features and apply sliding window detectors or deformable part-based models [11]. However, these techniques require a large number of labeled training examples, which makes them unsuitable for classes with sparse examples.

Other approaches explore fine-grained image classification [3,26] to capture discriminative image regions that distinguish between different classes. In [1], randomization and discrimination are combined in a computationally efficient scheme to achieve fine-grained classification in a large feature space.

Multi-label image classification [17,34,16] differs from multi-class recognition [28] in that a single image is classified using multiple labels. Multi-label classification usually incorporates modelling the correlation between the labels, which significantly boosts the semantic classification performance [16]. In [35], Genetic Algorithm optimization is utilized for filtering the selected features, which are then used for classification. Unlike us, they rely on multi-label training data.



(a) Sample training images from 'Food', 'Plants', and 'Healthcare' categories



(b) Sample testing images from our collected dataset

**Fig. 2.** Sample images from our collected dataset

There have been relatively limited research attempts to recognize products in images [27,30,15,20,33]. In systems like [15,20], a product image search engine is built. However, these systems deal with recognizing only one product per image, with training images having similar conditions to query images. In [27], cross-dataset single product recognition is targeted through query object segmentation combined with iterative retrieval. They achieve good results in searching an image that contains only a single product. In our work, however, we target multi-products image parsing.

A related dataset to our proposed one is presented in [22], however the dataset size is much smaller than ours with only 120 grocery products in the training set. Each product category represents a single specific product (i.e. no hierarchies or classes of products). In [27], a sports product image dataset is collected. Both the training images and 67 query images contain a single product per image. Other existing object datasets include Caltech 101 [9], Caltech 256 [13], and VOC [8] datasets, which target more general-purpose object category recognition. Fine-grained object datasets include Oxford Flower [23], and Stanford Dogs [18] dataset. We run our experiments on our proposed dataset, as well as the one presented in [22].

In the following sections, we first present our proposed novel dataset in section 3. Then, we detail our approach in section 4. Experimental results and performance evaluation are discussed in section 5. Finally, conclusions and future work are summarized in section 6.

## 3 Grocery Products Dataset

We built a new supermarket products dataset, which can be used in multi-label, fine-grained cross-dataset object recognition. Our dataset consists of 8350

training images spanning 80 product categories, downloaded from the Web. Each grocery product is represented by exactly one training image, taken in ideal studio conditions with a white background. On the other hand, test images are taken in real-life scenarios using a mobile phone. Each test image contains several products, ranging from 6 to 30 products per image. Test images are taken with different lighting conditions, viewing angles and zoom levels, introducing many challenges to the recognition process.

Training images are organized in hierarchical categories. For example, a Snickers chocolate bar is classified as "Food/Candy/Chocolate". The number of training images in each fine-grained category ranges from 25 to 415 images, with an average of 112 different retail products in each category – one training image for each product. We added an additional label for background regions. The images for the background label represent shelves and price tags, extracted from test images. Examples of training images are shown in figure 2 (a).

Grocery products introduce many challenges to the object recognition problem. First, many products have similar appearance, with only minor differences in the color of the package, size of the package, or some text on the box. Also, non-planar products, like bottles or jars, lower the matching performance, considering that we only have one training image per product. Furthermore, evaluation images contain very little background regions, which makes it a rather challenging task to recognize every single product in the image. This is due to the fact that other regions in the image represent confusing background clutter relative to the specific region of each object.

One of the main goals of this work is to investigate cross-dataset multi-label image classification. Accordingly, our evaluation set is collected in completely different conditions from the training set. A total of 680 images are taken in different grocery stores covering the different classes in the training dataset. Testing images impose additional challenges, like specularities, different viewing angles, rotated or occluded products as shown in figure 2 (b).

We ran our experiments on 27 classes of the "Food" category products in addition to the background class, which represents shelves and price tags, with a total number of 3235 images. Deformable objects, like nuts bags, chips and bakery are included in the "Food" category of our datast. To evaluate the performance of our algorithm, we annotated 680 test images with all the products from the 27 training classes. The ground truth of each test image specifies bounding boxes with a corresponding single product label for each bounding box. A single bounding box covers a group of instances of the same product in a test image as shown in figure 3.

## 4 Exemplar-Based Multi-label Image Classification

In this section, we describe the design and implementation details of our algorithm. Figure 1 shows an overview of our system. Our proposed technique consists of three main steps. The first two steps filter the best matching products to a given test image through two successive ranking procedures. The third



**Fig. 3.** Sample test images with ground truth annotations from our proposed dataset

step simultaneously localizes and infers the total number of objects present in the test image through globally minimizing an energy function.

Although we are going to use specific algorithms for each step of our pipeline, any other algorithm that fits to a single step can be applied. For example, we used discriminative decision trees for multi-class ranking, but it would suffice to use SVM or k-NN for classification. Similarly, we can use any matching algorithm for the second step of our pipeline like SIFT flow or sparse features matching.

### 4.1 Multi-class Ranking

To reduce our search space from thousands of possible matches to tens up to a few hundreds of images, we train a classification model using the given training images, and then use a voting scheme, explained below, to retrieve the top-ranked object classes.

For training, we use the recently proposed discriminative random forests [1] technique. The training set contains a single image for each product with a total number of 3235 images in 27 classes. We extract dense SIFT feature descriptors [21] on each image with a spacing of four pixels, with five patch sizes: 8, 12, 16, 24 and 30. Visual vocabulary codebook of 256 code words is then constructed using k-means. Descriptors are assigned to code words using Locality-constrained Linear Coding (LLC) [31].

To retrieve the top ranked object classes for a given test image, we designed a voting algorithm, which first divides the test image into grids with different sizes. We, then, classify each grid region separately using the trained model. We gather votes for each class in the trained model by counting the number of grid segments belonging to that class. For each test image, we return the top $k$ classes.

Our proposed class ranking technique handles two important challenges faced in cross-dataset object recognition, specifically in the products recognition domain. First, each object in the test image is surrounded by many other objects that have very similar features, which can easily confuse the classifier. By dividing the image into patches of different sizes, we limit such confusion. Second, by collecting the total number of votes for grids, we lower the impact of regions in the image suffering from difficult imaging conditions in affecting the final classification decision.

In the experiments section, we detail the parameters used for multi-ranking, and we show how the multi-label ranking performance is improved through gathering votes over grid patches rather than classifying the whole image once.

### 4.2 Fast Dense Pixel Matching

To achieve simultaneous recognition and localization of specific object instances in each test image, we apply fast dense pixel matching through deformable spatial pyramid matching [19]. No training is required to perform this step. Furthermore, it contributes to the scalability of our system, in such a way that adding new specific objects to the dataset does not require retraining the random forests step, as long as these objects fall under one of the pre-existing classes.

The goal is to rank the images in terms of appearance agreement while enforcing geometrical smoothness between neighboring pixels. The matching objective can be expressed formally by minimizing the energy function [19]:

$$E(t) = \sum_{i} D_i(t_i) + \alpha \sum_{i,j \in N} V_{ij}(t_i, t_j), \tag{1}$$

where $Di$ is a data term which measures the average distance between local descriptors within node $i$ in the first image to those located within a region in the second image after shifting by $t_i$. $V_{ij}$ is a smoothness term, $\alpha$ is a constant weight, and $N$ denotes pairs of nodes linked by graph edges. The energy function is minimized using loopy belief propagation. Training images are scaled to 200x200 pixels, and test images are scaled to 600x450 pixels. We use the mean difference of dense color SIFT feature descriptors of patch size of 4 as our data term. In all the experiments, the value of $\alpha$ was fixed at 0.005 following [19].

A segmentation mask is obtained specifying the inferred location of every pixel in each matched image with respect to the current test image. The matching costs, along with the segmentation masks are used in the next step of our pipeline to produce the final multi-labeling results as explained in section 4.3.

### 4.3 A Genetic Algorithm-Based Multi-label Image Classification

Once we obtain a ranked list of matching correspondences, we then consider only the top $N$ images, which will be in the range of very few tens of images, to obtain our final multi-labeling results. We formulate our problem in a genetic algorithm (GA) optimization model [12]. The quality of a given solution is determined using a fitness function, which is the objective function to be minimized using GA.

To define our multi-label image classification objective function, let $q$ be our current test image. We want to find the $L \subset N$ images that minimize the following energy function:

$$E(L) = \alpha \sum_{l \in L} D_{lq}(l, q) + \beta U_{Lq}(L, q) + \gamma C_L, \tag{2}$$

where $D_{lq}$ is the data term between image $l \in L$ and the current test image $q$, $U_{Lq}$ is the uncoverage term, which measures the proportion of pixels not covered




(a) (b)

**Fig. 4.** Sample (a) training and (b) testing images from Grozi-120 dataset

by any image $l$ in L when the whole set is warped to $q$. Finally, our context term $C_L$ models the prior knowledge about the co-occurrence of recognized products in the query image. $\alpha$, $\beta$, and $\gamma$ are weight parameters.

We chose our data term $D_{lq}$ to measure the mean difference between the dense SIFT descriptors between the two images, as defined by [19]. We experimented with adding other features like normalized RGB color histogram. However, the performance was worse with global color, where most products are colorful, and the lighting conditions of test images are very different from training images.

The coverage term $U_{Lq}$ penalizes results that do not cover a big proportion of the test image. If we define $S_{lq}$ to be the set of non-overlapping pixels in $q$ covered by $l$ when warped to $q$, then $U_{Lq}$ can be defined as:

$$U_{Lq}(L,q) = 1 - \frac{1}{z}\sum_{l \in L} |S_{lq}|, \tag{3}$$

where $z$ is the total number of pixels in the test image $q$, and $|S_{lq}|$ is the cardinality of the set $S_{lq}$. This, again, helps in overcoming the challenge of having multiple database images with very similar visual appearance. Such images will all be ranked as top matches, but for only one object in the test image. Just taking the top ranked results, would then yield very poor coverage of the objects present in the test image.

The context term $C_L$ models the prior probability that the labels which appear in the final retrieved set of images occur together. In other multi-label classification approaches, this knowledge is usually inferred from the training images. In our case, this knowledge cannot be obtained from the training images, as each image in our training set contains only a single product. We overcome this problem through utilizing the hierarchical structure of our solution. We model the prior distribution such that images (or labels) which fall under the same category are more likely to occur together than those which fall in different categories. The probability of co-occurrence is higher for more restrictive categories than for broader categories.

$$C_L = 1 - \sum_{l_i, l_j \in L} \tilde{P}(l_i, l_j), \tag{4}$$

where $\tilde{P}(l_i, l_j)$ is the prior distribution over the pairwise co-occurence of labels.

The overall energy function in Eq. 2 is globally minimized using constrained genetic algorithm (GA) [12]. We represent the population of possible solutions as a binary vector of length N, where each element represents the decision of inclusion for each image in the set. We used the "ga" method provided in the Matlab Global Optimization Toolbox. To constrain the type of children that the algorithm creates at each step to be binary, we implemented special creation, crossover, and mutation functions [6].

## 5 Experimental Results

**Datasets:** We evaluated the performance of our approach on two datasets:

1. 680 annotated test images from the proposed **"Grocery Products"** dataset, with a total number of 3235 products in 27 leaf node classes. Test images contain products of all subcategories in the "Food" category ranging from 6 to 30 product items per image. Regions in the test images which contain objects that do not belong to the database are given a null label.
2. 885 extracted test images from **GroZi-120** [22] dataset. There is a total of 676 training images representing 120 grocery products. Each product is represented by 2-14 training images with an average of 5.6 images. There are no classes of products (i.e. each class has only one specific product). The originally provided test images were unsuitable, since each image contains a single product item. No shelves images were provided. We, instead, extracted video frames from the provided 29 video files, each representing the whole frame as shown in figure 4. Each test image contains 4-15 grocery product items. Training images are downloaded from the web in ideal conditions, while test images are taken in grocery stores with different conditions.

**Implementation Details:** We trained 100 trees with a maximum depth of 10. We gathered votes for each test image over 57 patches of 5 different grid sizes. The motivation behind choosing these values is explained in section 5.4. The values of the parameters for the energy function (defined in Eq. 2): $\alpha$, $\beta$ and $\gamma$ are optimized using coordinate descent as detailed in section 5.2.

**Evaluation Metrics:** We measure the performance of our proposed system using three metrics: mean average precision (mAP), mean average product recall (mAPR), and mean average multi-label classification accuracy (mAMCA) [2]. We chose non-standard measures because standard measures usually address the performance of single-instance retrieval. mAP is measured by computing the average precision over all test images for different values of the number of top matched images ($n$) that we consider in the matching step, and then the mean is taken over all values of $n$ (ranging from 5 to 70). Averaging helps to capture the joint precision-recall performance. We count groups of specific products in a test image not individual product items (figure 3). We measure the mAPR by computing the average labeling performance (recall) of the retail product items present in an image, and then the mean is computed across all images. To

**Table 1.** Multi-label image classification performance for baseline labeling, different versions of our system, and state-of-the-art classification and retrieval techniques

| Method | mAP(%) | mAMCA(%) | mAPR(%) |
|---|---|---|---|
| Baseline [19] | 13.53 | 11.77 | 37.33 |
| Full | 23.49 | 21.19 | 43.13 |
| without global optimization | 16.93 | 15.07 | 43.36 |
| with ground truth ranking | 42.56 | 38.02 | 45.63 |
| ground truth ranking without global optimization | 30.7 | 27.8 | 68.5 |
| FV(1024 dim) | 8.62 | 6.41 | 20.73 |
| FV(4096 dim) | 11.26 | 9.95 | 22.14 |
| FV(4096 dim) + RANSAC | 12.3 | 10.1 | 24.5 |
| HE($k$=200000, $h_t$=22) | 4.26 | 3.96 | 12.13 |

compute mAMCA over the test dataset D, suppose $Y_x$ is the set of ground truth labels for test image $x$, and $P_x$ is the set of prediction labels. We can define the multi-label score for image $x$ as $score(P_x) = \frac{|Y_x \cap P_x|}{|Y_x \cup P_x|}$, and

$$accuracy_D = \frac{1}{|D|} \sum_{x \in D} score(P_x), \tag{5}$$

To analyze the performance of our multi-class ranking approach, we, also, use two measures: mean average recall (mAR) per-class and mean average accuracy (mAA) over the test images. We vary $K$, i.e. the number of predicted classes from 1 to the total number of classes, and measure the true positive and false positive rates accordingly.

In the next sections, we first perform quantitative and qualitative evaluation of our system (Sec. 5.1). We perform in-depth analysis of our GA optimization in Sec. 5.2. Results on the GroZi-120 dataset are reported in Sec. 5.3. Finally, multi-class ranking and runtime efficiency are discussed in Sec. 5.4 and Sec. 5.5.

### 5.1 Multi-label Image Classification Performance

To evaluate the performance of our proposed approach, we vary the number of predicted classes of the multi-class ranking ($K$) from 1 to the total number of classes and report the mAMCA and mAPR values on different variants of our system (see table 1): (1) full system (Full), (2) our system without performing global optimization (i.e. retreive all the $n$ top-ranked images from the dense pixel matching results on the $k$ top-ranked class categories), (3) our system if we have perfect ranking performance of the multi-class ranking step, and (4) ground truth ranking without performing global optimization. We compare the performance of our algorithm to state-of-the-art classification and instance-level image retrieval techniques, Fisher Vectors [25] (FV) and Hamming Embeddings [14] (HE). For FV, we use 1024 and 4096 dim. encodings without PCA. We, also, compare to FV (4096 dim) with Geometric Consistency Checks with RANSAC




**Fig. 5.** Examples of two multi-label classification results. Left column shows the test image, then the retrieved products, and finally their inferred locations in the test image.

re-ranking on the top 100 images. For HE, we use $k = 200,000$ visual words for building the bag-of-words histogram representation which was shown to yield good performance and we use a fixed Hamming threshold $h_t = 22$ following [14]. Finally, we compare to the baseline method of ranking all the images by just dense pixel matching score [19] and taking the top $n$ matches.

Our full system achieves 23.49% mAP and 21.19% mAMCA over all the 680 test images, which outperforms the baseline method by over 9%. Our method also significantly performs better than other state-of-the-art approaches. FV and HE are efficient algorithms which achieve impressive precision on other benchmarks. However, for our case, the distribution of the training data from which the GMM model is built (for FV), or the BOF dictionary is built (for HE) is significantly different from the data distribution of the test set. In addition, these methods are better suited for general rather than fine-grained object recognition.

We also show the performance results if we run our dense pixel matching ranking and global optimization steps using the images of the ground truth classes (i.e., we assume that the multi-label ranking step gives a perfect ranking of predicted categories for each test image). This yields a substantial improvement in the mAP and mAMCA, which shows that our system's performance could be further improved by experimenting with different classification techniques. When evaluating the system, the parameters are optimized for maximizing the precision and accuracy of recognition. Accordingly, the recall performance is not much improved given the chosen values of the parameters. Showing the improvement of precision for the same achieved recall values gives an indication of how ground truth ranking can improve the performance of the system.

To verify the impact of our global optimization step, we report results when we pick the top $n$-ranked images from the matching step as our final multi-label classification result without any global optimization. We have two cases: (1) with random forests model ranking, and (2) with ground truth ranking. We notice that the mAMCA degrades by more than 6%, as more irrelevant images appear in the final result. In case of ground truth ranking, our system still performs better with global optimization.

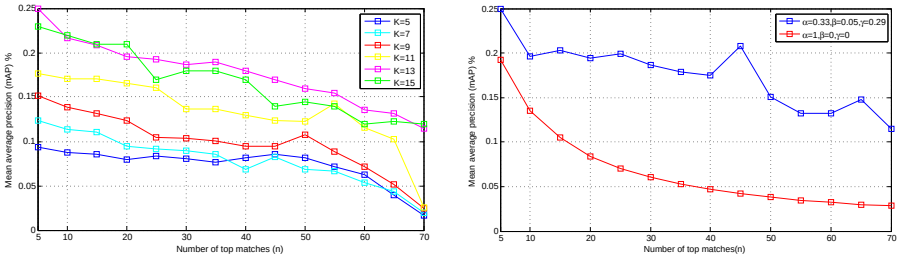



**Fig. 6.** (a) Mean average precision as a function of the total number of matches ($n$) for different values of the number of filtered classes ($K$). (b) Mean average precision as a function of the total number of top matches ($n$) when turning on (and off) the GA optimization. Our GA step significantly yields better performance.

In figure 5, we show sample results from running our full system on different test images to illustrate the effectiveness of our proposed technique. We show the original test image, the inferred labels, and their predicted locations in the test image. Failure cases are mainly due to significant visual resemblance between training images (like the cereal box in the figure), severe specularities, and blurry conditions of test images. We also fail to recognize wrong facing products, which can be addressed with additional training images.

**Discussion:** Although the absolute value of the results may seem unsatisfactory, we have to consider the challenging settings of our problem. Other fine-grained classification datasets like Caltech-UCSD Birds dataset [32] report a state-of-the-art performance of average accuracy of 19.2%, considering that there are 15 training images per category, and each image contains only one object instance. Whereas for our dataset, we have a single training image per category, and each test image contains an average of 20 objects per image.

### 5.2 GA Optimization

We analyze the performance of our GA optimization by investigating the mAP when choosing different parameter values for $K$, $n$, $\alpha$, $\beta$ and $\gamma$. We first study the effect of the number of filtered classes $K$, in the multi-ranking step, and the number of top matches $n$, in the dense pixel matching step, on the mAP performance of the system. Figure 6 (a) shows the mAP as a function of $n$ for different values of $K = 5, 7, 9, 13, 15$. For each combination of $n$ and $K$, we obtain the optimal values of $\alpha$, $\beta$ and $\gamma$ which maximize the mAP using coordinate descent optimization. It is shown that increasing the number of classes $K$ generally improves the mAP. However, as $K$ keeps increasing, more noise is added to the filtered set which decreases the mAP. Best performance is obtained for $K = 13$ classes. As expected, mAP decreases as $n$ increases, but at the same time the recall improves. In figure 6 (b), we plot the mAP as a function of $n$ for

**Table 2.** Performance on Grozi-120. System parameters are optimized to maximize average precision rate.

| Method | mAP | mAMCA | mAPR |
|---|---|---|---|
| Baseline [19] | 7.62 | 6.24 | 16.59 |
| Full | 13.21 | 7.5 | 9.37 |
| without global optimization | 9.54 | 7.1 | 17.56 |
| with ground truth ranking | N/A | 43.03 | 43.03 |
| FV(1024 dim) | 4.44 | 5.49 | 12.50 |
| FV(4096 dim) | 7.34 | 5.74 | 15.16 |
| FV(4096 dim) + RANSAC | 8.13 | 6.65 | 15.2 |
| HE($k$=200000, $h_t$=22) | 6.32 | 5.23 | 10.54 |

$K = 13$ when turning off the GA optimization, by setting $\alpha = 0$, $\beta = 0$ and $\gamma = 0$, and when turning on the GA optimization by fixing $\alpha = 0.33$, $\beta = 0.05$ and $\gamma = 0.29$ (obtained using coordinate descent). Our GA step significantly improves the mAP performance. Also, our curve is flatter which shows that our method is more tolerant to noise imposed by adding more images in the dense pixel matching step.

### 5.3 Performance on GroZi-120 Dataset

We ran our experiments on 885 extracted test images (see figure 4). We used the same metrics and compared to the same approaches as in Sec. 5.1. Our system significantly outperforms other methods and the baseline method as shown in table 2. Please note that mAP value for ground truth ranking variant of our system will always have a value of 100.0% because each product category represents a specific product in Grozi-120 dataset. For similar reasons, the ground truth ranking without global optimization setting cannot be applied to Grozi-120 dataset. Figure 7 shows sample results from running our algorithm on Grozi-120. We effectively recognize and infer the locations of the objects in a test image.

We note that our system achieves lower mAP values on the Grozi-120 dataset than on our proposed dataset. This is due to the fact that there are only 5.6 images per product (which represents a class) on average for training which greatly degrades the results of the discriminative random forests. This is verified in the significant improvement of the system performance when using ground truth ranking. Further more, a large proportion of the test images in the Grozi-120 dataset suffer from blurriness. Nevertheless, our system outperforms other approaches. Also, there is no available prior information. We adjusted the prior model to be the $l1$-norm of the total number of recognized products in the image.

### 5.4 Multi-class Ranking Analysis

To demonstrate the impact of our multi-class ranking scheme, we report the mAA and mAR values using (1) different number of segments (i.e. votes), as



**Fig. 7.** Examples of two multi-label classification results on Grozi-120. Left column shows the test image, then the retrieved products, and finally their inferred locations.

**Table 3.** Multi-class ranking analysis. Baseline is the binary classification of images.

| | Baseline | 1 seg | 5 seg | 57 seg |
|---|---|---|---|---|
| mAA | 25.56 | 63.55 | 62.52 | **64.00** |
| mAR | 22.4 | 57.22 | 53.32 | **58.35** |

opposed to (2) using the whole image (i.e. 1 segment) for ranking, and (3) performing binary classification of a test image (baseline). We have experimented with different, empirically chosen, segment sizes. Results in table 3 show that ranking classes through gathering classification votes consistently yields better performance. The impact of regions in the image that suffer from specularities or very wide variation in viewing angles is regularized by considering other patches that have better conditions.

### 5.5 Runtime Efficiency

Our system consists of 3 steps: (1) Multi-class ranking, (2) fast dense pixel matching, and (3) global optimization. We ran our experiments on a single 2.4G CPU with 4 GB of RAM without code optimization. Step (1) takes an average of 0.2 seconds per test image, not considering feature extraction time. Step (2) takes 0.35 seconds per each matching operation, and finally step (3) converges to an optimal solution in around 1.4 seconds when we consider the top 20 images for optimization. Accordingly, the total runtime of our algorithm is 1.95 seconds, where the time for dense pixel matching is parallelized for $n$ top-ranked images. The most time consuming task is the LLC feature extraction.

## 6 Conclusions and Future Work

We presented a fast and scalable novel approach to recognize and localize all specific product instances in a retail store image with minimum training. We perform

cross-dataset multi-label image classification, where each label is represented by just one instance in our training set. We also propose a new large-scale retail products dataset with thousands of different labels. Experiments showed that our system significantly yields better results than existing state-of-the-art classification and instance-level retrieval methods on both our proposed dataset, and Grozi-120 dataset. Although we apply our proposed method to the grocery products use case, our algorithm is general and can be applied to other multi-label image classification problems. Accordingly, the next step for us is to experiment with applying our system in other domains and compare the performance with previous methods on available benchmarks.

## References

1. Bangpeng, Y., Aditya, K., Fei-Fei, L.: Combining randomization and discrimination for fine-grained image categorization. In: CVPR (2011)
2. Boutella, M.R., Luob, J., Shena, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recognition 37(9), 1755–1771 (2004)
3. Branson, S., Wah, C., Schroff, F., Babenko, B., Welinder, P., Perona, P., Belongie, S.: Visual recognition with humans in the loop. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 438–451. Springer, Heidelberg (2010)
4. Csurka, G., Dance, C., Bray, C., Fan, L.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision (2004)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
6. Deep, K., Singh, K.P., Kansal, M.L., Mohan, C.: A real coded genetic algorithm for solving integer and mixed integer optimization problems. Applied Mathematics and Computation 212(2), 505–518 (2009)
7. Duan, G., Huang, C., Ai, H., Lao, S.: Boosting associated pairing comparison features for pedestrian detection. In: ICCV Workshop on Visual Surveillance (2009)
8. Everingham, M., Gool, L.V., Williams, C., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV 88(2) (2010)
9. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In: Workshop on Generative-Model Based Vision, CVPR (2004)
10. Fei-Fei, L., Fergus, R., Torralba, A.: Recognizing and learning object categories. In: ICCV Tutorial (2005)
11. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE T. Pattern Anal. 32(9), 1627–1645 (2010)
12. Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley (1989)
13. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Technical report, Caltech (2007)
14. Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 304–317. Springer, Heidelberg (2008)

15. Jing, Y., Baluja, S.: Pagerank for product image search. In: WWW (2008)
16. Jurie, Y.S.F.: Improving image classifcation using semantic attributes. IJCV 100(1), 59–77 (2012)
17. Kang, F., Jin, R., Sukthankar, R.: Correlated label propagation with application to multi-label learning. In: CVPR (2006)
18. Khosla, A., Jayadevaprakash, N., Yao, B., Fei-Fei, L.: Novel dataset for fine-grained image categorization. In: First Workshop on Fine-Grained Visual Categorization, CVPR (2011)
19. Kim, J., Liu, C., Sha, F., Grauman, K.: Deformable spatial pyramid matching for fast dense correspondences. In: CVPR (2013)
20. Lin, X., Gokturk, B., Sumengen, B., Vu, D.: Visual search engine for product images. In: Multimedia Content Access: Algorithms and Systems II (2008)
21. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60(2), 91–110 (2004)
22. Merler, M., Galleguillos, C., Belongie, S.: Recognizing groceries in situ using in vitro training data. In: CVPR (2007)
23. Nilsback, M.E., Zisserman, A.: A visual vocabulary for ower classification. In: CVPR (2006)
24. Pandey, M., Lazebnik, S.: Scene recognition and weakly supervised object localization with deformable part-based models. In: ICCV (2011)
25. Perronnin, F., Liu, Y., Sanchez, J., Poirier, H.: Large-scale image retrieval with compressed fisher vectors. In: CVPR (2010)
26. Sharma, G., Jurie, F., Schmid, C.: Discriminative spatial saliency for image classification. In: CVPR (2012)
27. Shen, X., Lin, Z., Brandt, J., Wu, Y.: Mobile product image search by automatic query object extraction. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 114–127. Springer, Heidelberg (2012)
28. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I, LNCS, vol. 3951, pp. 1–15. Springer, Heidelberg (2006)
29. Torralba, A., Efros, A.: Unbiased look at dataset bias. In: CVPR (2011)
30. Tsai, S.S., Chen, D.M., Chandrasekhar, V., Takacs, G., Cheung, N.M., Vedantham, R., Grzeszczuk, R., Girod, B.: Mobile product recognition. In: ACM Multimedia (ACM MM) (2010)
31. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: CVPR (2010)
32. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-ucsd birds 200. Technical report cns-tr-201, Caltech (2010)
33. Winlock, T., Christiansen, E., Belongie, S.: Toward real-time grocery detection for the visually impaired. In: CVAVI (2010)
34. Zha, Z., Hua, X., Mei, T., Wang, J., Qi, G., Wang, Z.: Joint multi-label multi-instance learning for image classification. In: CVPR (2008)
35. Zhang, M., Pena, J., Robles, V.: Feature selection for multi-label naive bayes classification. Information Sciences 179(19), 3218–3229 (2009)

**Appendix KEJRIWAL01**


Page 76 of 169

```
001    9921412433001451
leader 03417cam a2200625 i 4500
005    20240220213017.0
006    m     o  d
007    cr mn|||||||||
008    150929t20152015njua    ob    100 0 eng d
050    #4$aTJ210.3 $b.I344 2015
037    ##$b00066573
019    ##$a956657992 $a1127214376
040    ##$aSTF $beng $erda $epn $cSTF $dMYG $dOCLCO $dOCLCF $dOCLCO $dIAS $dOCLCA $dCEF $dOCLCQ $dCUS $dOCLCO $dSFB $dCOO $dREB $dVT2 $dWYU $dLVT $dUMC $dOCLCO $dOCLCQ
020    ##$z9781479987573
020    ##$z1479987573
020    ##$a9781479987580
020    ##$a1479987581
024    7#$a10.1109/TePRA34669.2015 $2doi
049    ##$aMAIN
111    2#$aIEEE International Conference on Technologies for Practical Robot Applications $n(7th : $d2015 : $cWoburn, Mass.)
264    #1$aPiscataway, NJ : $bInstitute of Electrical and Electronics Engineers, $c[2015]
264    #4$c©2015
245    10$a2015 IEEE International Conference on Technologies for Practical Robot Applications (TePRA) : $bCrowne Plaza Boston-Woburn 15 Middlesex Canal Park Woburn, Massachusetts, 01801 USA, May 11-12, 2013.
246    1#$iAvailable from some providers with title: $aTechnologies for Practical Robot Applications (TePRA), 2015 IEEE International Conference on
246    1#$iAlso known as: $aTePRA 2015
336    ##$atext $btxt $2rdacontent
337    ##$acomputer $bc $2rdamedia
338    ##$aonline resource $bcr $2rdacarrier
300    ##$a1 online resource : $billustrations (some color)
347    ##$atext file $bPDF $2rda
520    8#$aWith an emphasis on practical applications and industrial participation, TePRA provides researchers, innovators, entrepreneurs and robotics professionals with a balanced forum in robotics science, engineering and technology
500    ##$aConference date in title should read 2015.
500    ##$a"IEEE Catalog number CFP15TEP-ART."
588    0#$aOnline resource; title from PDF title page (IEEE Xplore, viewed on September 29, 2015).
504    ##$aIncludes bibliographical references.
650    #0$aRobotics $vCongresses.
650    #0$aRobots, Industrial $vCongresses.
650    #0$aMobile robots $vCongresses.
650    #6$aRobotique $vCongrès.
650    #6$aRobots industriels $vCongrès.
650    #6$aRobots mobiles $vCongrès.
650    #7$aMobile robots $2fast
650    #7$aRobotics $2fast
650    #7$aRobots, Industrial $2fast
655    #7$aproceedings (reports) $2aat
655    #7$aConference papers and proceedings $2fast
655    #7$aConference papers and proceedings. $2lcgft
655    #7$aActes de congrès. $2rvmgf
710    2#$aInstitute of Electrical and Electronics Engineers, $eissuing body.
776    0#$z1479987581
856    40$3IEEE Xplore $uhttps://ieeexplore.ieee.org/servlet/opac?punumber=7172788
994    ##$a92 $bOCACL
035    ##$a(OCoLC)922537699 $z(OCoLC)956657992 $z(OCoLC)1127214376
035    ##$a(OCoLC)ocn922537699
```

TOP
SEND TO
VIEW ONLINE
ITEM DETAILS
LINKS

CONFERENCE PROCEEDING

# 2015 IEEE International Conference on Technologies for Practical Robot Applications (TePRA) : Crowne Plaza Boston-Woburn 15 Middlesex Canal Park Woburn, Massachusetts, 01801 USA, May 11-12, 2013.

IEEE International Conference on Technologies for Practical Robot Applications (7th : 2015 : Woburn, Mass.)

2015; Piscataway, NJ : Institute of Electrical and Electronics Engineers

Online access >

## Send to

E-MAIL | PERMALINK | CITATION | PRINT | ENDNOTE | EASYBIB | EXPORT RIS | BIBTEX | EXPORT TO EXCEL

## View Online

**Full text availability**

IEEE Xplore conference proceedings [SHOW LICENSE]
*Access for all on campus; login required from off-campus.*
*Details about who can access online resources*

IEEE Xplore conference proceedings [SHOW LICENSE]
*Access for all on campus; login required from off-campus.*
*Details about who can access online resources*

## Item Details

| | |
|---|---|
| **Title** | 2015 IEEE International Conference on Technologies for Practical Robot Applications (TePRA) : Crowne Plaza Boston-Woburn 15 Middlesex Canal Park Woburn, Massachusetts, 01801 USA, May 11-12, 2013. |
| **Summary** | With an emphasis on practical applications and industrial participation, TePRA provides researchers, innovators, entrepreneurs and robotics professionals with a balanced forum in robotics science, engineering and technology The presentations selected through IEEE s rigorous peer review process highlights new uses of algorithms, components, and whole systems in practical robot applications While papers are solicited in all related areas of robotics and automation, the conference tracks include?Medical and Rehabilitation?Military and Security?Actuation and Manipulation?Consumer and Industrial?Best Practices and Experience Reports?Robust and Reliable Robotics Technologies?Resource Aware Algorithms?Power & Communications?Sensing?Software Development Tools. |
| **Author** | IEEE International Conference on Technologies for Practical Robot Applications (7th : 2015 : Woburn, Mass.) ><br>Institute of Electrical and Electronics Engineers, issuing body. > |
| **Subjects** | Robotics -- Congresses ><br>Robots, Industrial -- Congresses ><br>Mobile robots -- Congresses ><br>Mobile robots ><br>Robotics ><br>Robots, Industrial > |
| **Genre** | proceedings (reports) ><br>Conference papers and proceedings > |
| **Language** | English |
| **Format** | 1 online resource : illustrations (some color) |
| **Technical Details** | Encoding format: PDF |
| **Alternate Title** | Available from some providers with title: Technologies for Practical Robot Applications (TePRA), 2015 IEEE International Conference on<br>Also known as: TePRA 2015 |

| | |
|---|---|
| **Description** | Conference date in title should read 2015.<br>"IEEE Catalog number CFP15TEP-ART."<br>Includes bibliographical references. |
| **Publisher** | Piscataway, NJ : Institute of Electrical and Electronics Engineers |
| **Publication Date** | 2015 |
| **Identifier** | ISBN: 9781479987580<br>ISBN: 1479987581<br>ISBN: 1479987581 (Additional Physical Format) |
| **OCLC Number** | 922537699 |
| **Alma IZ Record Number** | 99161913815501452 |
| **Alma NZ Record Number** | 99321412433001451 |

## Links

Staff View

**Appendix KUMAR01**

```
001     99320262437101451
leader  03393cam a2200613 i 4500
005     20240220213017.0
006     m     o  d
007     cr cn|||||||||
008     141010t20142014njua    ob    100 0 eng d
050     #4$aTJ210.3 $b.I344 2014
019     ##$a972629655 $a974263976 $a1026449144 $a1127203765
040     ##$aCOO $beng $erda $epn $cCOO $dIEXCP $dCUS $dOCLCF $dMYG $dOCLCO $dOCL $dOCLCO $dIAS $dREB $dOCLCA $dCEF $dOCLCQ $dLVT $dOCLCO $dSFB $dUMC $dOCLCQ
020     ##$z9781479946068
020     ##$z1479946060
020     ##$z9781479946044
020     ##$z1479946044
020     ##$z9781479946051
020     ##$z1479946052
024     7#$a10.1109/TePRA33408.2014 $2doi
049     ##$aMAIN
111     2#$aIEEE International Conference on Technologies for Practical Robot Applications $d(2014 : $cWoburn, Mass.)
264     #1$aPiscataway, NJ : $bIEEE, $c[2014]
264     #4$c©2014
245     10$a2014 IEEE Conference on Technologies for Practical Robot Applications (TePRA).
246     1#$iAlso known as: $aTePRA 2014
246     1#$iAvailable from some providers with title: $aTechnologies for Practical Robot Applications (TePRA), 2014 IEEE International Conference on
246     3#$a2014 IEEE International Conference on Technologies for Practical Robot Applications (TePRA)
336     ##$atext $btxt $2rdacontent
337     ##$acomputer $bc $2rdamedia
338     ##$aonline resource $bcr $2rdacarrier
300     ##$a1 online resource (various pagings) : $billustrations
520     8#$aAnnotation $bWith an emphasis on practical applications and industrial participation, TePRA provides researchers, innovators, entrepreneurs and robotics
500     ##$aConference held at Woburn, MA on 14-15 April 2014.
500     ##$a"IEEE Catalog Number CFP14TEP-CDR."
588     0#$aOnline resource; title from PDF copyright page (IEEE Xplore, viewed November 21, 2014).
504     ##$aIncludes bibliographical references.
650     #0$aRobotics $vCongresses.
650     #0$aRobots, Industrial $vCongresses.
650     #0$aMobile robots $vCongresses.
650     #6$aRobotique $vCongrès.
650     #6$aRobots industriels $vCongrès.
650     #6$aRobots mobiles $vCongrès.
650     #7$aMobile robots. $2fast $0(OCoLC)fst01024301
650     #7$aRobotics. $2fast $0(OCoLC)fst01098997
650     #7$aRobots, Industrial. $2fast $0(OCoLC)fst01099080
655     #7$aConference papers and proceedings. $2fast $0(OCoLC)fst01423772
710     2#$aInstitute of Electrical and Electronics Engineers, $eissuing body.
776     0#$z1479946044
776     0#$z1479946060
856     40$3IEEE Xplore $uhttps://ieeexplore.ieee.org/servlet/opac?punumber=6861557
994     ##$a92 $bOCACL
035     ##$a(OCoLC)893192305 $z(OCoLC)972629655 $z(OCoLC)974263976 $z(OCoLC)1026449144 $z(OCoLC)1127203765
035     ##$a(OCoLC)ocn893192305
```

CONFERENCE PROCEEDING

# 2014 IEEE Conference on Technologies for Practical Robot Applications (TePRA).

IEEE International Conference on Technologies for Practical Robot Applications (2014 : Woburn, Mass.)

2014; Piscataway, NJ : IEEE

Online access

TOP
SEND TO
VIEW ONLINE
ITEM DETAILS
LINKS

## Send to

E-MAIL | PERMALINK | CITATION | PRINT | ENDNOTE | EASYBIB | EXPORT RIS | BIBTEX | EXPORT TO EXCEL

## View Online

**Full text availability**

IEEE Xplore conference proceedings [SHOW LICENSE]

*Access for all on campus; login required from off-campus.*

*Details about who can access online resources*

## Item Details

| | |
|---|---|
| **Title** | 2014 IEEE Conference on Technologies for Practical Robot Applications (TePRA). |
| **Summary** | Annotation With an emphasis on practical applications and industrial participation, TePRA provides researchers, innovators, entrepreneurs and robotics professionals with a balanced forum in robotics science, engineering and technology The presentations selected through IEEE s rigorous peer review process highlights new uses of algorithms, components, and whole systems in practical robot applications While papers are solicited in all related areas of robotics and automation, the conference tracks include Medical and Rehabilitation Military and Security Actuation and Manipulation Consumer and Industrial Best Practices and Experience Reports Robust and Reliable Robotics Technologies Resource Aware Algorithms Power & Communications Sensing Software Development Tools. |
| **Author** | IEEE International Conference on Technologies for Practical Robot Applications (2014 : Woburn, Mass.)<br>Institute of Electrical and Electronics Engineers, issuing body. |
| **Subjects** | Robotics -- Congresses<br>Robots, Industrial -- Congresses<br>Mobile robots -- Congresses<br>Mobile robots<br>Robotics<br>Robots, Industrial |
| **Genre** | Conference papers and proceedings |
| **Language** | English |
| **Format** | 1 online resource (various pagings) : illustrations |
| **Alternate Title** | Also known as: TePRA 2014<br>Available from some providers with title: Technologies for Practical Robot Applications (TePRA), 2014 IEEE International Conference on |
| **Description** | Conference held at Woburn, MA on 14-15 April 2014.<br>"IEEE Catalog Number CFP14TEP-CDR."<br>Includes bibliographical references. |
| **Publisher** | Piscataway, NJ : IEEE |
| **Publication Date** | 2014 |
| **Identifier** | ISBN: 1479946044 (Additional Physical Format)<br>ISBN: 1479946060 (Additional Physical Format) |
| **OCLC Number** | 893192305 |
| **Alma IZ Record** | 99161855365801452 |


Number
Alma NZ Record Number 9932026243710145 1
Links
Staff View

**Appendix MERLER01**

# MARC View

| | | | |
|---|---|---|---|
| LEADER 01055nam a22002412 4500 | | | |
| 001 39330052 | | | |
| 003 SIRSI | | | |
| 005 20221114154312.0 | | | |
| 006 m d | | | |
| 007 cr n | | | |
| 008 070226s2007 nju s\|\|\|\|\|\|\| 2\|eng\|d | | | |
| 020 | | | a\| 9781424411795 |
| 020 | | | a\| 1424411793 (Trade Paper) c\| USD 518.00 Retail Price (Publisher) 9\| Active Record |
| 037 | | | a\| CH37898 b\| 00066573 |
| 040 | | | a\| BIP US d\| WaSeSS |
| 245 | 0 | 0 | a\| 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) h\| [electronic resource] |
| 260 | | | a\| Piscataway : b\| IEEE c\| Jan. 2007 |
| 506 | | | a\| License restrictions may limit access. |
| 521 | | | a\| Scholarly & Professional b\| IEEE |
| 655 | | 0 | a\| Electronic books. |
| 856 | 4 | 0 | u\| http://SK8ES4MC2L.search.serialssolutions.com/?sid=sersol&SS_jc=TC0000818058&title=2007%20IEEE%20Conference%20on%20Computer%20Vision%20and%20Pattern%20Recognition%20%28CVPR%29 |
| 955 | | | a\| Books In Print, (c) 2012 R.R. Bowker LLC |
| 949 | | | a\| Electronic resource w\| ASIS c\| 1 i\| 39330052-1001 l\| ONLINE m\| ONLINE r\| Y s\| Y t\| ONLINE u\| 11/17/2022 o\| .STAFF. 360 MARC |

ASK A LIBRARIAN

PennState

Copyright @2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

Legal Statements | PSU Hotlines

PENN STATE UNIVERSITY LIBRARIES

Libraries Home (814) 865-6368
Libraries Intranet (Staff Only)
Accessibility Help
Website Feedback
Policies and Guidelines
Staff Directory

CONNECT WITH PENN STATE UNIVERSITY LIBRARIES

Facebook
Twitter
Instagram

# 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [electronic resource]

Published: Piscataway : IEEE Jan. 2007

Access Online

serialssolutions.com

ASK A LIBRARIAN

**Restrictions on Access:**
License restrictions may limit access.

**Genre(s):**
Electronic books

**ISBN:**
9781424411795
1424411793 (Trade Paper)

**Audience Notes:**
Scholarly & Professional IEEE

View MARC record | catkey: 39330052

PennState

Copyright @2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

Legal Statements | PSU Hotlines

PENN STATE UNIVERSITY LIBRARIES
Libraries Home (814) 865-6368
Libraries Intranet (Staff Only)
Accessibility Help
Website Feedback
Policies and Guidelines
Staff Directory

CONNECT WITH PENN STATE UNIVERSITY LIBRARIES
Facebook
Twitter
Instagram

IEEE.org IEEE *Xplore* IEEE SA IEEE Spectrum More Sites Donate Cart Create Account Personal Sign In

Browse My Settings Help

Access provided by: Penn State University Sign Out

Access provided by: Penn State University Sign Out

All

ADVANCED SEARCH

Conferences > 2007 IEEE Conference on Compu...

# Recognizing Groceries in situ Using in vitro Training Data

**Publisher: IEEE**

Cite This PDF

Michele Merler ; Carolina Galleguillos ; Serge Belongie **All Authors**

85 Cites in Papers

4 Cites in Patents

709 Full Text Views

Alerts

Manage Content Alerts

Add to Citation Alerts

Abstract

Document Sections

1. Introduction
2. Related work
3. GroZi-120 database
4. Object detection and recognition algorithms
5. Experiments and Results

Show Full Outline

Authors

Figures

References

Citations

Keywords

Metrics

Downl PDF

**Abstract:**
The problem of using pictures of objects captured under ideal imaging conditions (here referred to as in vitro) to recognize objects in natural environments (in situ) is ... **View more**

**Metadata**

**Abstract:**
The problem of using pictures of objects captured under ideal imaging conditions (here referred to as in vitro) to recognize objects in natural environments (in situ) is an emerging area of interest in computer vision and pattern recognition. Examples of tasks in this vein include assistive vision systems for the blind and object recognition for mobile robots; the proliferation of image databases on the web is bound to lead to more examples in the near future. Despite its importance, there is still a need for a freely available database to facilitate study of this kind of training/testing dichotomy. In this work one of our contributions is a new multimedia database of 120 grocery products, GroZi-120. For every product, two different recordings are available: in vitro images extracted from the web, and in situ images extracted from camcorder video collected inside a grocery store. As an additional contribution, we present the results of applying three commonly used object recognition/detection algorithms (color histogram matching, SIFT matching, and boosted Haar-like features) to the dataset. Finally, we analyze the successes and failures of these algorithms against product type and imaging conditions, both in terms of recognition rate and localization accuracy, in order to suggest ways forward for further research in this domain.

**Published in:** 2007 IEEE Conference on Computer Vision and Pattern Recognition

More Like This

Footnotes

**Date of Conference:** 17-22 June 2007

**DOI:** 10.1109/CVPR.2007.383486

**Date Added to IEEE *Xplore*:** 16 July 2007

**Publisher:** IEEE

**ISBN Information:**

**Conference Location:** Minneapolis, MN, USA

**Print ISSN:** 1063-6919

Contents

## SECTION 1.
# Introduction

Object detection and recognition are important tasks in computer vision. Nearly all of the algorithms that address these problems demand large amounts of labeled training examples for each object. Moreover, it is generally necessary that the training examples be acquired under environmental conditions (illumination, pose, etc.) that are representative of what will be observed in the testing data.

In many real-world applications, however, having access to training data for which the appearance is drawn from the same distribution as that of the testing data is not practical. Examples of such applications include assistive vision technology for the blind and object recognition for mobile robots. Consider for example the Semantic Robot Vision Challenge (SRVC)[1], a new research competition designed to push the state-of-the-art in image understanding and automatic acquisition of knowledge from large unstructured databases of images (such as those generally found on the web). Central to the challenge is the problem of training a "laptop-on-wheels" object recognizer using only a textual list of objects (provided shortly before the competition) and access to an internet image search engine. In the case of our application of interest, we seek to design a portable vision system for helping blind people find groceries in a supermarket, with a starting point again consisting of a textual list — in this case a shopping list — and access to publicly available grocery images on the internet.

The web contains an immense collection of structured and unstructured image databases and offers the potential to generate useful models for image recognition and localization. In order to explore this potential in at least one application domain, we built a multimedia database consisting of 120 different grocery items easily retrievable from the web. For every product, two different recordings are available: *in vitro* images extracted from the web (captured under ideal imaging conditions), and *in situ* images extracted from camcorder video collected inside a grocery store. We refer to a grocery product with its corresponding UPC code as an object. Each product has several different image examples that were extracted from the web and from video captures. The intent of this dataset is to serve as a seed upon which the set of images can grow dynamically by user interaction in the future. In particular, methods such as [5] could be used to discover and label large amounts of *in situ* object examples in photos that currently lack textual annotation using *in vitro* training data from online resources such as Amazon Groceries.

Our goal in this paper is to provide a highly varied real world multimedia database for studying object recognition and detection with the above-described distinction between training and testing data, as well as baseline performance ∼ prybski/SRVCI of several approaches used to solve object localization and recognition. In Section 2 we first review the state of the art object detection and recognition algorithms used in our baseline study, and then we discuss existing databases of common use in the computer vision community. In Section 3 we introduce the GroZi-120 database, which is publicly available, with a detailed description of what it contains and how it was created. Section 4 presents the features used and the different models for localization and recognition. Section 5 shows the results on the database and finally in Section 6 we discuss the results and propose ideas for future work.

## SECTION 2.
# Related work

### 2.1. Detection/recognition algorithms

Color histogram matching is one of the first algorithms ever applied to detect and recognize objects in images and videos. Swain and Ballard's early work on color object recognition by means of fast matching color histograms by intersecting them [20] opened the way to many different approaches with a common ground. Computational complexity, which has always been one major bottleneck of the histogram extraction and comparison based search tasks, has recently been overcome with the introduction of the integral histogram by Fatih Porikli [19]. On the other hand, the choice of which colorspace or chrominance plane to use is still an open issue [14].

In addition to color features, one can employ methods based on on shape [4] and/or gray-scale object descriptors. There are two distinct problems linked to such approaches: interest point detection and description of the distribution of smaller-scale features within the interest point neighborhood. The most commonly used interest point detector is probably the Harris corner detector [12], while for the second problem, David Lowe's SIFT [16] descriptor has been shown to outperform the others, as it is invariant to scale and rotation transformations. This property can be explained by the fact that it captures a substantial amount of information about the spatial intensity patterns, while at the same time being robust to small deformations or localization errors.

In the framework of object detection, a major contribution has been offered by Viola and Jones [21], who introduced a fast and reliable classifier. They suggested using a cascade of classifiers, each of which is a combination of a subset of weak learners selected via Adaboost [8]. The weak learners use Haar-like features, which are responses of filters computed extremely quickly by employing integral images.

### 2.2. Computer vision databases

Over the years many datasets have been introduced in the computer vision community, in order to provide a mean of evaluation for the different algorithms developed. Here we present a sample of recent ones, summarizing their characteristics. In this way we will make evident the innovative contribution offered by the GroZi-120 dataset, which we will introduce in Section 3

The Pascal Object Recognition Database Collection [1] consists of data gathered for an object recognition in natural scenes challenge that took place in 2005 and 2006. In its final version, the collection contains 5,304 images, provided by Microsoft Research Cambridge and collected from the photo-sharing web-site Flickr, of 10 object classes. All images are annotated with instances of all the classes, for a total of 9,507 labelled objects. The set presents variability of scale, pose, background clutter and degree of occlusion for every object.

Caltech 101, now expanded into Caltech 256 [7], contains 30607 images, grouped into 256 object classes, with a mean of 119 images per class. It also includes a special category for clutter and background. It is widely used for object recognition but is not recommended for object localization.

Another example is SOIL-47 [13], a database of household objects, many of the same shape, viewed over a significant portion of the viewing sphere. The images show mainly multicolored objects, many of them consisting of planar surfaces (boxes) and with generally complex color structure. The database contains 24 objects with approximately planar surfaces and 22 complex scenes. Both objects and scenes are presented against a black background, in absence of clutter. Three different kinds of appearance variation are included: 3D viewpoint, illumination intensity and occlusion/distractors. A database like SOIL-47 may represent an extremely rich *in vitro* training set for the problem we investigate, but it lacks the *in situ* counterpart that is needed for testing.

A more recent computer vision database is the Amsterdam Library of Object Images (ALOI) [9], a collection of one-thousand small objects. The creators of the dataset systematically varied viewing angle, illumination angle, and illumination color for each object, and additionally captured wide-

baseline stereo images. It includes over a hundred images per object, yielding a total of 110,250 images. Again, no clutter or complex background are present.

Finally, the ETH-80 [15] database consists of 80 objects from 8 chosen categories captured in high-resolution color images, with segmentation masks provided for every image. Each object is represented by 41 images from viewpoints spaced equally over the upper viewing hemisphere.

Apart from the Caltech 256 and the PASCAL dataset, all the reported databases present a uniform one-color background, so objects are easily segmentable from the background (which is usually black and presents no clutter). Moreover, in almost all cases no occlusion is present and the sizes and position of objects inside the images are normalized, as pointed out in [18]. In contrast, we propose a dataset that contains images that present a significant difference in quality (between training and testing) and where individual objects have both cluttered and uniform background. Therefore it can be used for object localization as well as recognition approaches, and it can be used for training and testing purposes, either with images or video captures.

## SECTION 3.
# GroZi-120 database

The GroZi-120[2] is a multimedia database of 120 grocery products. It was created in order to address the problem of learning using training data that differs in quality from the testing data for object recognition and localization tasks. The objects belonging to it vary in color, size, opacity, shape and rigidity, as can be seen in Figure 2. The dataset introduces variabilities not systematically offered in previous available collections. In fact, many of our testing images contain multiple instances of the same object, which may present partial occlusion and truncation, as well as size and orientation variations. Furthermore, the location of the product varies considerably from image to image and different objects are found in the same frames. All these properties are exemplified in Figure 4. Every product has two different representations in the database: one captured *in vitro* and another *in situ*. The *in vitro* images are isolated and captured under ideal imaging conditions (e.g., stock photography studio or a lab) and they can be found on the web, more specifically in grocery web stores such as Froogle [3].

In order to obtain *in vitro* representations we queried the web using a list of approximately 4000 UPC codes associated with the products of a local grocery store. A Perl script was generated to automatically crawl the web looking for images of the products by using the UPC code. In case the system did not respond positively to a query, a textual description of the product was obtained from the UPC online database[4], and the query was performed again with the new keywords. A manual selection of the downloaded instances was executed in order to eliminate duplicates and images representing products different from the desired one. In order to make the *in vitro* images usable as a training set for any algorithm, we set their background to transparent and we obtained a binary mask in order to extract only the useful information out of every image, as shown in Figure 5(a).

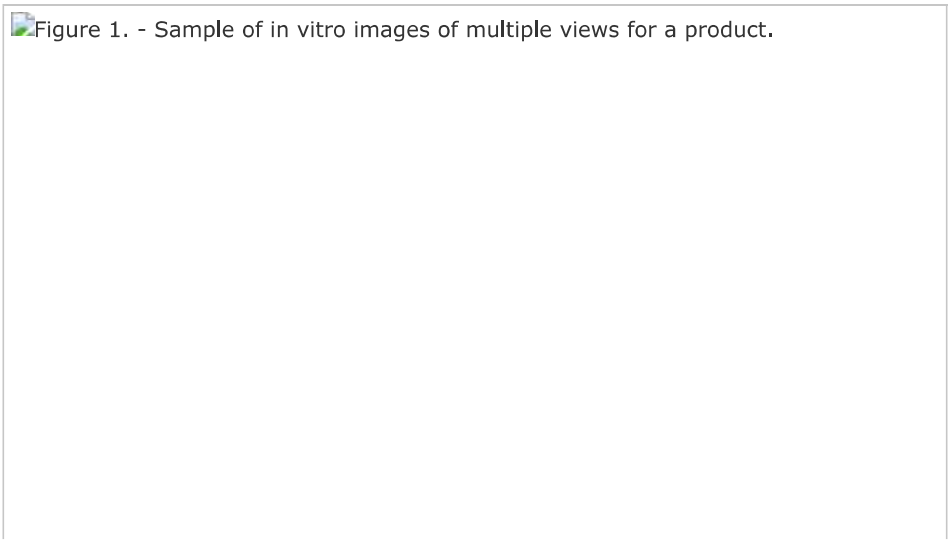


**Figure 1.** Sample of in vitro images of multiple views for a product.

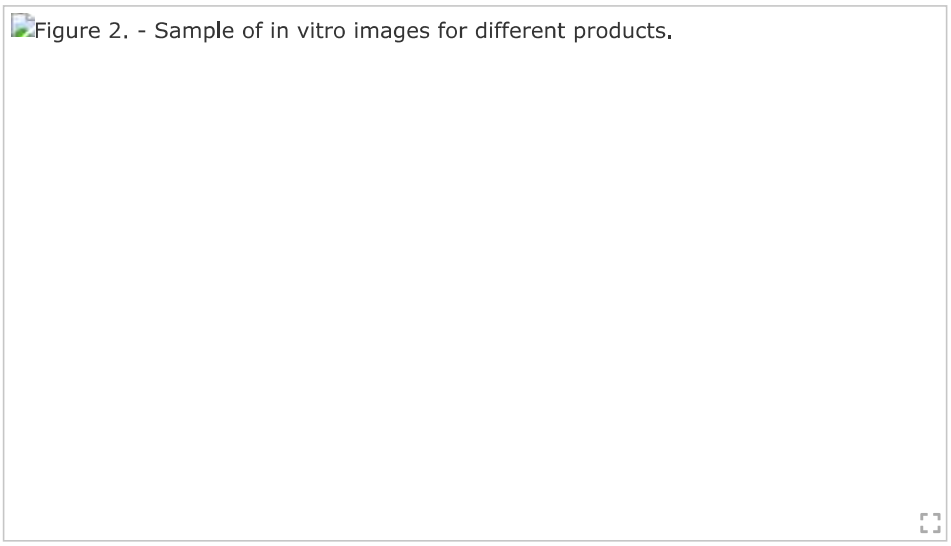



**Figure 2.** Sample of in vitro images for different products.

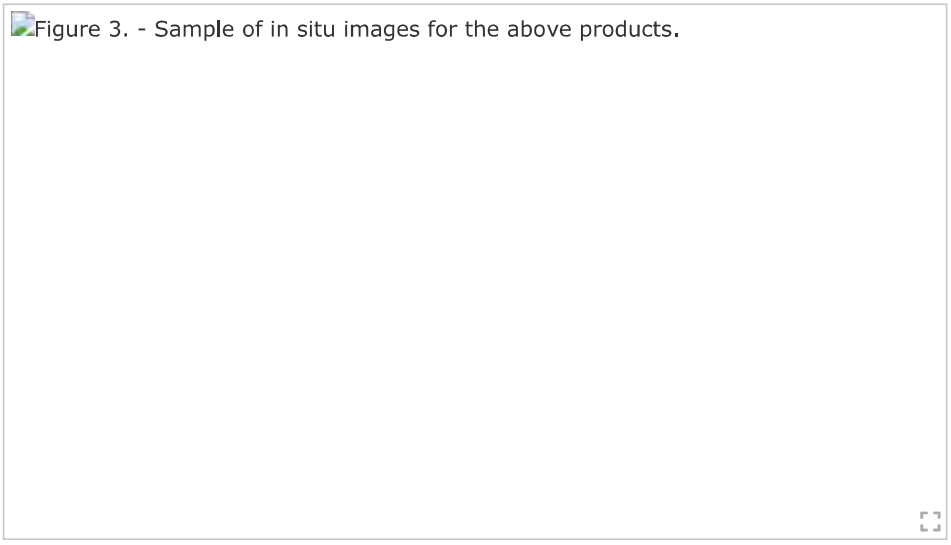
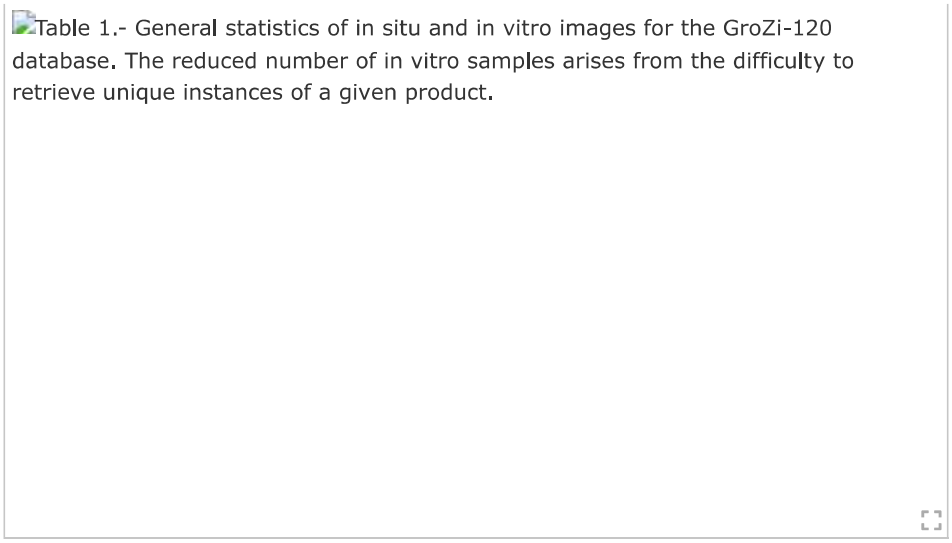



**Figure 3.** Sample of in situ images for the above products.

**Table 1.** General statistics of in situ and in vitro images for the GroZi-120 database. The reduced number of in vitro samples arises from the difficulty to retrieve unique instances of a given product.


Table 1.- General statistics of in situ and in vitro images for the GroZi-120 database. The reduced number of in vitro samples arises from the difficulty to retrieve unique instances of a given product.

Therefore the *in vitro* images become easy to analyze and at the same time, coming from different online vendors and stock photo suppliers, they include a variety of illuminations, sizes and poses. Figure 1 shows the different views of a particular product represented by *in vitro* images. In contrast, *in situ* representations come from natural environments (real world). Figure 3 shows the *in situ* representations of the objects in Figure 2. We went to the grocery store that provided us the UPC list (a campus convenience store), and shot 29 videos on the same day at 30 fps, encoded as Divx 5.2.1 with a bitrate of 2000 kbps using a VGA resolution MiniDV camcorder, for a total of around 30 minutes of footage. Such videos include every product present in the *in vitro* part of the dataset. These images were selected every 5 frames and were stored together with their location in the video (video number, frame number, rectangle coordinates), as shown in Figure 5(b). These images present variations in scale, illumination, reflectance, color, pose and rotation, while the video frames provide a cluttered background. One of the benefits of this data is that it represents the typical low quality of a real world image. Hence, different algorithms for object recognition (where *in vitro* images can be compared to *in situ* images) and object localization (search for products in videos) can be tested on this dataset. Table 1 reports statistics about the dataset. Grozi-120 constitutes a database of mixed multimedia content, with images and videos representing the same objects under different visual conditions, together with their text annotations. The creation of such a dataset involved different multimedia resources, spanning from the web to digital camcorders.

## SECTION 4.
# Object detection and recognition algorithms

In order to provide a baseline characterization of the level of difficulty inherent to this problem domain, we tested a selection of popular object detection and recognition approaches and studied their performance on the GroZi-120 dataset. Below we describe the different features used and the approaches implemented in detail.

### 4.1. Features and dissimilarity measures

**Color Histogram**

We first tested our database with chrominance planes belonging to 3 different color spaces: YCbCr, HSV and Lab. A preliminary study showed that the ab plane from Lab provided the best results. We computed histograms of 16 bins per channel, a and b, calculated separately, for a total of 32 bins. Then we generated a histogram of the a and b channels for every *in vitro* image in the dataset. Histograms belonging to the same product were subsequently averaged, bin per bin, in order to obtain a final template histogram representative of the object.

**SIFT**

We computed SIFT [16] keypoints for every in *vitro* image in the data set in order to represent images using scale and rotation invariant descriptors. The keypoints were computed using binaries provided by the UCLA Vision Labs [5] Each grocery product was represented by a "bag of keypoints" extracted from the *in vitro* images corresponding to a particular object. Therefore the calculated keypoints correspond to the different views of the same object without including background. The background pixels were set to zero when the masks were applied to the images. The product shape information is implicitly (albeit weakly) captured by the descriptors, since they are computed on the masked images.

**Boosted Haar-like features**

We used the implementation in the Haar training utility of the Intel OpenCV library[6] samples were synthetically created from the *in vitro* images by applying randomly generated perspective distortions, until 200 positive samples were obtained, including the *in vitro* instances. The dimensions of such images were obtained by computing the average ratio of all the masked *in vitro* images and then resizing them to have a longest dimension of 50 pixels.

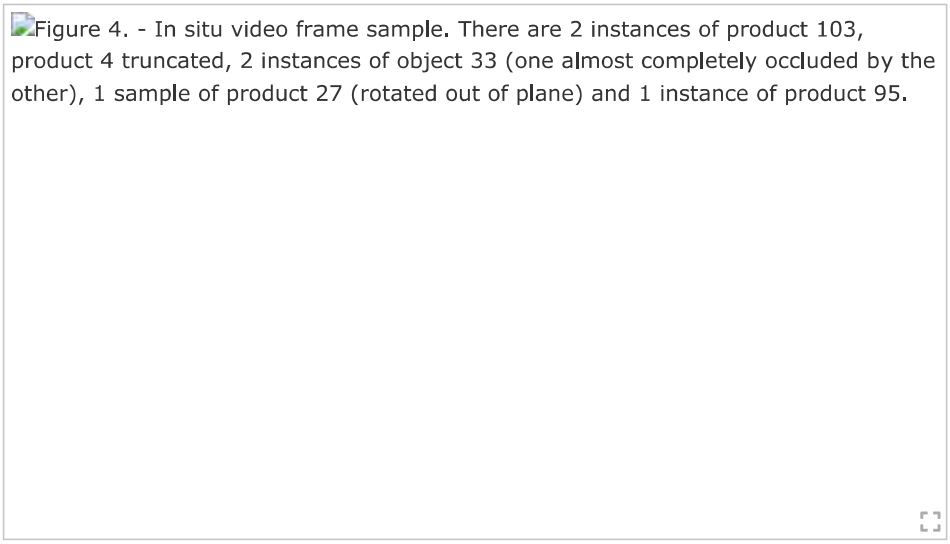



**Figure 4.** In situ video frame sample. There are 2 instances of product 103, product 4 truncated, 2 instances of object 33 (one almost completely occluded by the other), 1 sample of product 27 (rotated out of plane) and 1 instance of product 95.

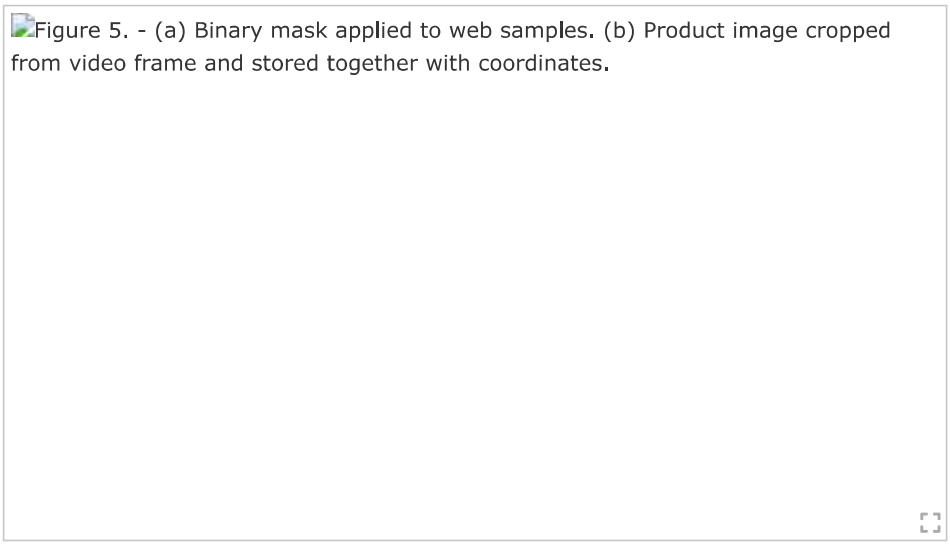



**Figure 5.** (a) Binary mask applied to web samples. (b) Product image cropped from video frame and stored together with coordinates.

### 4.2. Recognition

In our recognition study, *in vitro* images were used as training data, while *in situ* images were used as test data. In color histogram matching we computed color histograms for every *in situ* image in the dataset. Then the distances between the *in vitro* image template and *in situ* image histograms were calculated, according to 3 different metrics: Euclidean, $\chi^2$ and histogram intersection ($L_1$

distance). Once the distances were computed, we calculated the ROC curves shown in Section 5 by integrating a bidimensional histogram of the distances. We obtained the best performance using color histogram intersection.

Recognition using SIFT proceeded similarly to the approach used for color histogram matching. For each product we computed the bag of features obtained from its *in vitro* images. Then we matched the features with the keypoints of every *in situ* image in the dataset. The distances between samples were represented by the number of matches between the *in vitro* and the *in situ* one. The ROC curves were computed in the same manner as for the color histogram matching.

With regard to training in Adaboost, the cascade was allowed to be 14 stages deep, with a maximum false alarms rate (FAR) of 0.5 per stage. The ROC curves were obtained by computing the distances to the *in situ* images resized to the dimensions of the training set of the particular product. Such distances were computed for the classifier by taking the weighted sums of the responses of the features selected at every stage and finding their difference from the thresholds of the stage. All the differences were then summed together to obtain a unique value for the distance.

### 4.3. Localization

Localization experiments were conducted by trying to identify the location of products present in video sequences, using *in vitro* images as the training set. Testing was performed for every product using 14 frames out of the videos containing the product, with its locations manually identified as ground truth, and 100 frames from the same videos not containing any of the objects in the dataset. We took into account the possible presence of multiple instances of the product in the same frame.

We used two different metrics to evaluate localization: a yes/no rule from which we obtain true positives and false positives rates, and a metric based on the average object area recall and the average detected area precision as defined in [17], that gives us the *overall recall* and *overall precision* rates. The first is a frame-based metric where a *yes* is given if the center of the detected box (meaning the best match) lies within the ground truth region and a *no* otherwise, as presented in [3]. In the second, the recall for an object is defined by the authors as the proportion of its area that is covered by the algorithm's output boxes for every frame and the overall recall is computed as the weighted average recall of all frames. Precision of an output box is defined as the proportion of its area that covers the ground truth objects and overall precision is the weighted average precision of all frames.

The color histogram matching approach relies on the integral histogram computation as in [19]. We compute the *in vitro* template histogram in the same way as in recognition. Then, we performed a frame by frame analysis first computing the integral histogram (for the a and b channels) of the whole frame and then moving a window at 5 different scales in raster scan order around the frame, computing the color histograms of different regions. Then those histograms are intersected with the *in vitro* template. If the distance is accepted by the system, it is stored to be later compared to all the other best matches in the frame. Finally a maximum of 6 windows are kept, which correspond to the best scores.

The SIFT approach for object localization consisted first in computing the bag of keypoints for every *in vitro* object and for each frame as in recognition. Since the *in vitro* image sizes are different, we normalized the coordinates of all the keypoints. Then, we matched the keypoints against the frame features as in [16]. In order to reduce the number of outliers and locate the object we centered a circle in the *in vitro* average image, with a diameter equal to the average image diagonal. Iterating over the matches we kept the circle containing the maximum number of matches. Subsequently we computed the centroid of the locations of the corresponding matching features in the frame, and also their average distance with respect to the centroid. Using the previous pair of matches, we found the ones inside a new circle with a diameter equal to the average distance in the *in vitro* and frame instances. If the number of frame matches found is greater or equal than the *in vitro* ones found we consider the object as detected, otherwise we use the matches that were not taken into account in the first place and proceed in the same way. If in both cases the condition is not satisfied, we do not consider the object as detected. This approach was performed for different circles so we could handle multiple instances of a product on a frame.

In the Adaboost based method, the classifier obtained from the set of boosted Haar-like features during the training phase is used to decide whether a series of rectangles analyzed at different scales

within each frame contains the product of interest. This process is performed by comparing the responses of the filters selected for every stage of the classifier during the training phase to thresholds also selected during the training.

## SECTION 5.
# Experiments and Results

In this Section we present the results for recognition and localization of the algorithms presented in the previous sections. For recognition, we chose 10 *in vitro* image samples for every product for training, corresponding to the highest number of SIFT keypoints, considering that a higher number of keypoints often means an image with lower blurriness. Synthetic images were created when the original number of samples was less than 10. Subsequently we chose 10 *in situ* images per object for testing using the same criteria, but not considering synthetic images. In Figure 6 we report the average performances. It can be noticed how the overall best performing algorithm is SIFT, while the color histogram matching curve (CHM), after an initial fiat behavior, rapidly grows towards a 99 percent match rate. Finally, Adaboost offers the worst performance. One of the reasons for such a difference relies on the lack of a consistent number of original samples for training, alleviated by the generation of synthetic images, which do not offer effectively diverse characteristics. SIFT and CHM, on the other hand, need only very few samples to extract the necessary information to perform recognition. Figures 7 and 8 present two interesting cases. In the case of product 1, there is a clear disparity between the performances of the different algorithms. While the SIFT performed well, probably due to the distinctiveness of the text and symbols on the product's box, the color histogram curve is affected by the predominance of the white color, which is not particularly distinguishable in the chrominance plane, and performs poorly since it is based on a more global approach. On the other hand, in Figure 8, CHM benefits from the multimodal histogram obtained from a product presenting colors (yellow, orange and green) very well distinguishable in the ab plane, while SIFT and Haar-like features encounter a lack of distinctive feature points on most of the product's surface.

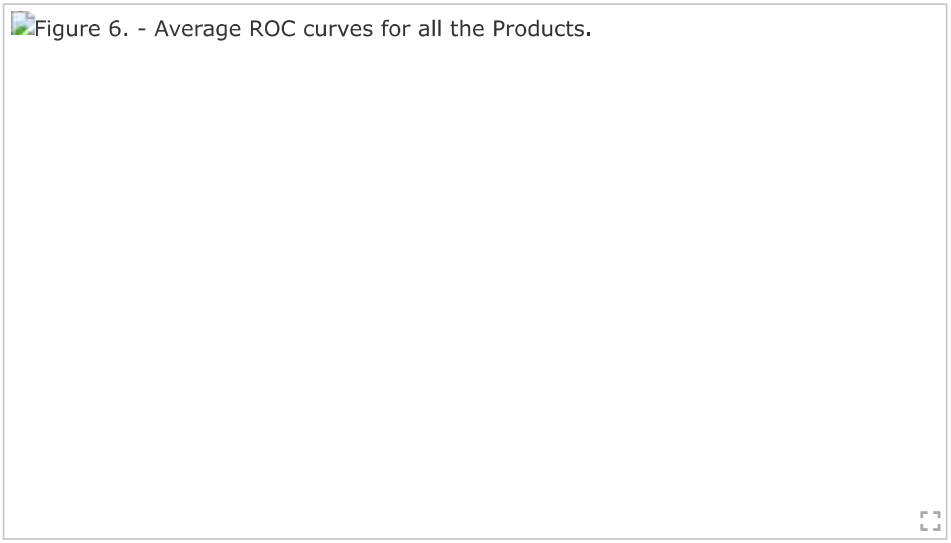



**Figure 6.** Average ROC curves for all the Products.

In the case of localization we used 100 frames with no instances of any product in the database to test the False Positive rates. Table 2 shows the localization rates in percentages for the three different algorithms. We use a yes/no rule to obtain true and false positives, based on whether the center of the detected box lies or not within ground truth region. Recall is defined as the proportion of the ground truth area covered by the detection boxes. Precision is defined as the proportion of the detection boxes that covers the ground truth area. The best and worst cases are selected considering the combination of all the 4 rates. Figure 9 presents the best and worst cases. Looking the results we can notice that SIFT outperforms the other methods, followed by CHM. For Adaboost the discussion reported for recognition still applies. The high values of the standard deviations in the table show that the methods do not perform uniformly over the set of products. This fact is due to

the complexity and variety of the dataset, which presents challenges of variable difficulty for any algorithm. Furthermore from the experiments emerges a discrepancy between the efficacy of the three methods in recognition and localization. While recognition yields decent results, the average performances in localization are poor.

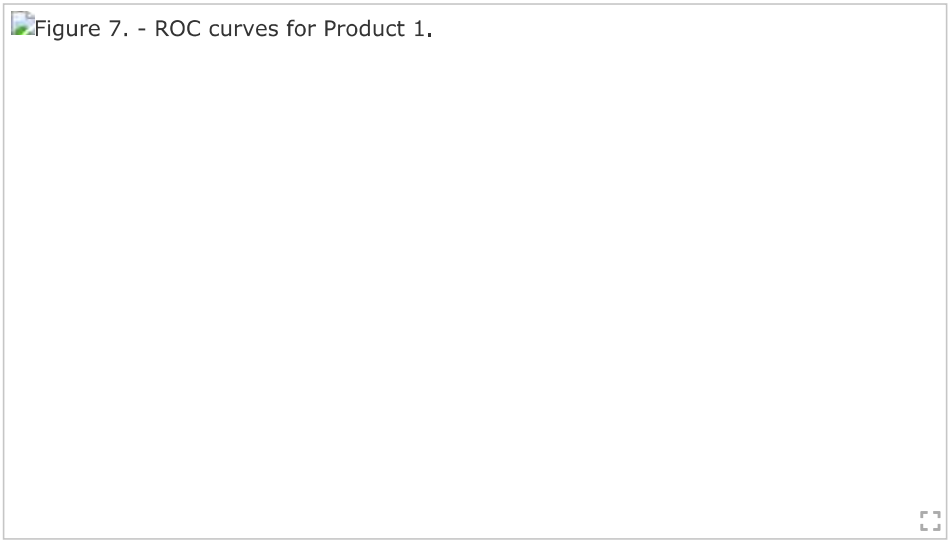



**Figure 7.** ROC curves for Product 1.

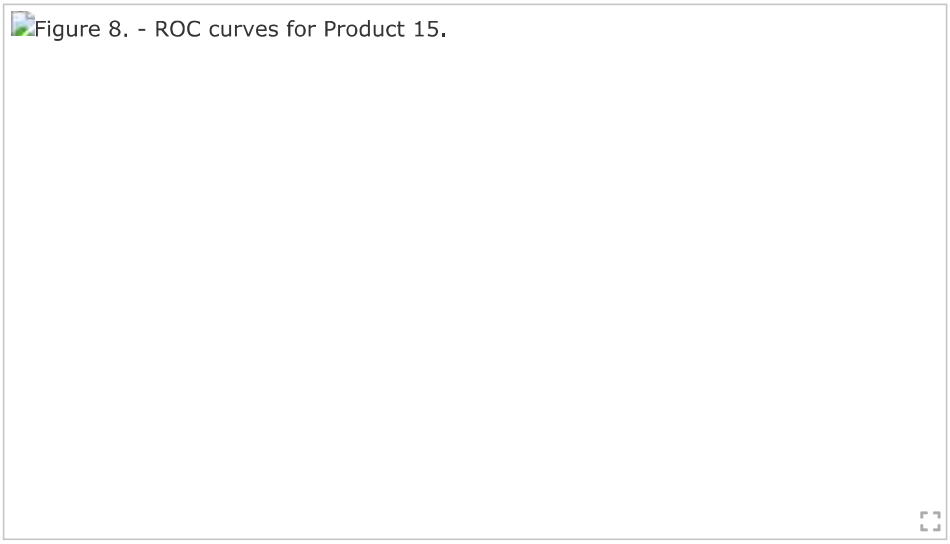



**Figure 8.** ROC curves for Product 15.

Such a discrepancy is due to the different nature of the tasks: in recognition we are still operating in a controlled environment, where although it is true that the samples we are comparing come from the different *in vitro* and *in situ* worlds, it is also true that those instances are segmented from the cluttered background, and therefore easier to analyze. On the other hand, localization must consider the full background, which contributes a significant amount of confusion to the problem. In this sense localization expresses the core of the problem: find the correspondence between *in vitro* samples and cluttered, noise corrupted, realistic *in situ* scenes. Figures 10, 11 and 12 are examples of different cases of the localization performance of the 3 algorithms on the dataset. From Figure 10 we see CHM perform well for product 52, while SIFT and boosted Haar-like features are not as good. In fact, object 52 has a distinctive orange color, which clearly stands out of the background. On the other hand, the remaining two methods cannot rely on such precious information and they are misled by the text on the neighboring products. Figure 11 shows good localization by all the methods on the product where SIFT performs best. The characteristics of the product (clear and distinctive colors, multiple pattern variations) generate informative features for every algorithm. Only boosted Haar like features are partially misled by a neighbor. Taking a closer look at the misdetection and a fraction of an *in vitro* sample, the similarity between the two is evident.

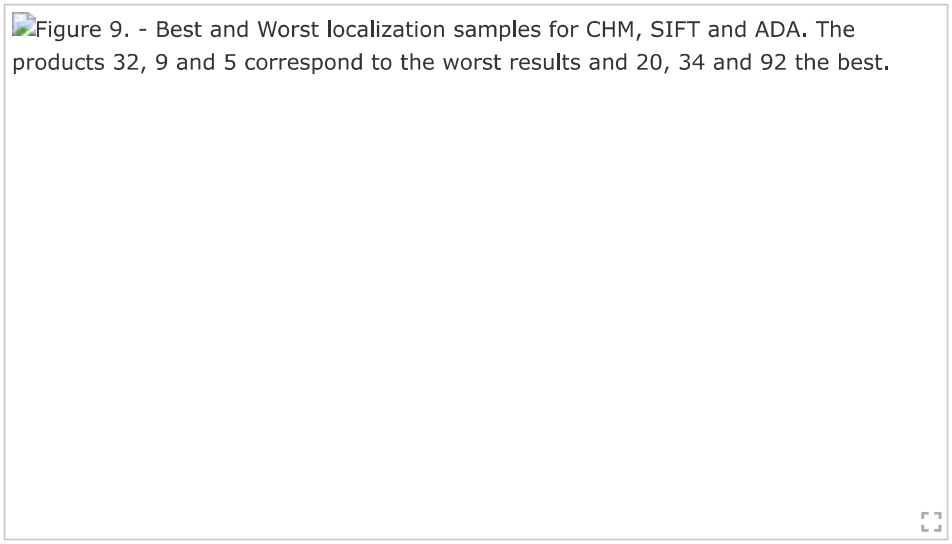
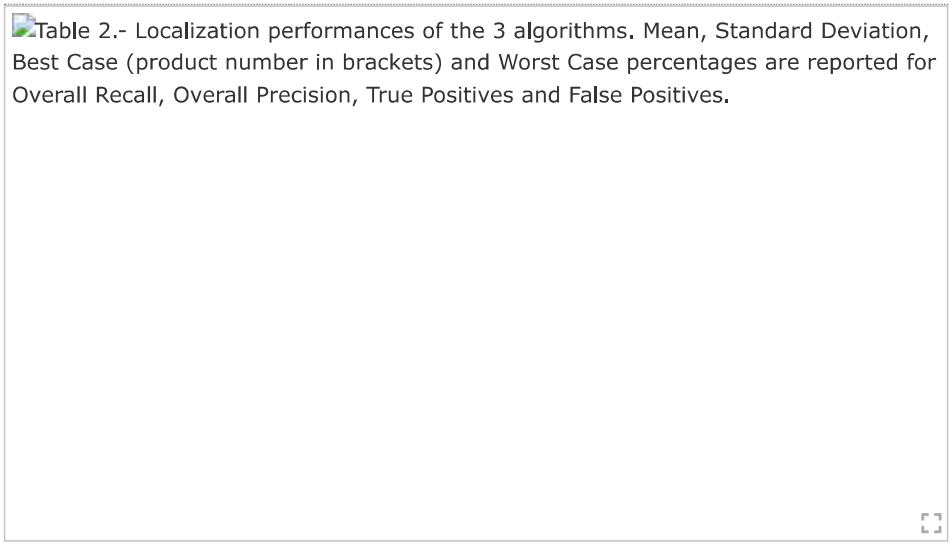



**Figure 9.** Best and Worst localization samples for CHM, SIFT and ADA. The products 32, 9 and 5 correspond to the worst results and 20, 34 and 92 the best.

**Table 2.** Localization performances of the 3 algorithms. Mean, Standard Deviation, Best Case (product number in brackets) and Worst Case percentages are reported for Overall Recall, Overall Precision, True Positives and False Positives.




Finally we present a bad localization example: product 74. CHM is misled by portions of the frame very similar in color to some parts of the *in vitro* image. For SIFT, the image is too rich with keypoints for the algorithm to be able to succeed, in particular because the product of interest is mainly composed by uniform patterns, so that the keypoints are found on corners in the writing. This together with the reflections on the palstic surface of the Yoohoo bottle prove deceptive for the algorithm. Boosted Haar like features on the other hand, finds (in grayscale) a product which is dark on the top and the bottom, and light in the middle, with some texture inside.

## SECTION 6.
# Discussion and future work

Our contributions in this paper include (a) a new multimedia database — GroZi-120 — for studying object recognition *in situ* (i.e., sitting in the real world) using training images from an *in vitro* source (i.e., captured under ideal conditions) and (b) baseline performance figures of three widely used recognition/detection algorithms that highlight the challenge presented by this database. The

Grozi-120 database contains both *in situ* and *in vitro* representations of the same products, and it presents a wide range of variability for each product with respect to size, color, rigidity, shape, illumination, viewpoint and quality.

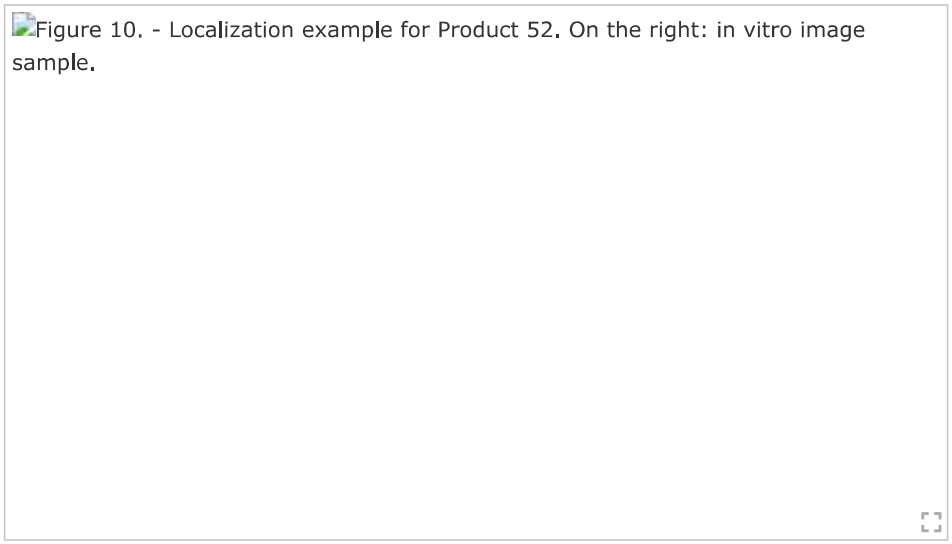



**Figure 10.** Localization example for Product 52. On the right: in vitro image sample.

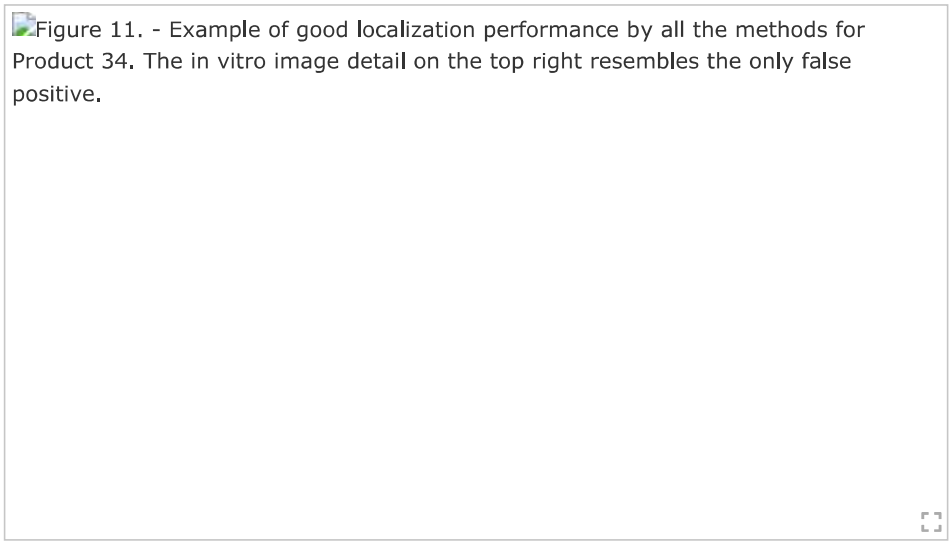



**Figure 11.** Example of good localization performance by all the methods for Product 34. The in vitro image detail on the top right resembles the only false positive.

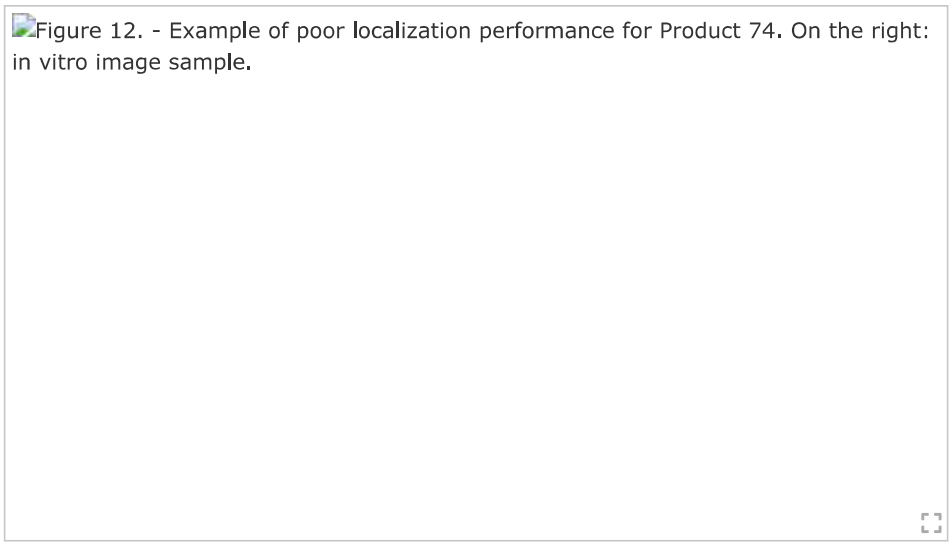


**Figure 12.** Example of poor localization performance for Product 74. On the right: in vitro image sample.

Gathering useful training information from images captured in ideal conditions is linked to the semantic web image retrieval issue, addressed among others by [22]. In this work the authors demonstrate the application of common image search metrics applied to images captured with a camera-equipped mobile device to find matching images on the Web or other general-purpose databases. The results of the testbed suggest the need for more precise and elaborate detection/recognition algorithms to solve the proposed problem. In fact, the aim of our work was to provide a baseline against which to test new methods. Color histogram matching may benefit from the works on color invariants, e.g., [11] and [10]. Another interesting direction may consist in a combination of the grayscale and color based methods, such as [2] and [6]. Our database offers not only a testbed for such approaches trying to link real world data to clean web or studio images, it can also be seen as an intermediate step or a bridge between the two representations of the same object. In fact, web and multimedia retrieval can benefit from a set of pre-existing labeled samples, while at the same time dynamically increasing and improving such a set. This type of problem can find applications in mobile robot navigation like the Semantic Robot Vision Challenge (SRVC) and also in assistive technologies for the visually impaired. In fact, in the case of a blind or visually impaired person that uses a device that recognizes products in a grocery store, it would be impractical to acquire *in situ* data every time we need to train the system, thus the *in vitro* data captured from the web is a good source of training data.

Therefore we intend as future work to use the *in vitro* data set as a seed to build upon, as the user base continues to use the database and expanding the *in situ* part as a means of evaluating different algorithms. Furthermore, we plan to include more objects in the database and expand the number examples per object. We also plan to fuse recognition and detection methods in order to improve the results as well as to make use of context information about physical object proximity, identifying products nearby on the shelf to improve localization of objects in natural scenes.

## ACKNOWLEDGMENTS

The authors would like to thank Vincent Rabaud, Stephan Steinbach, Anton Escobedo, Al Labotski, Robert Meza and Michael McQuen for valuable help and feedback, as well as all the people involved in the GroZi project at UCSD. Michele Merler was supported by the California Institute for Telecommunications and Information Technology (Calit2) 2006 Summer Undergraduate Research Scholarship Program. Partial support was also provided by NSF CAREER Grant #0448615 and the Alfred P. Sloan Research Fellowship.

Authors
Figures
References
Citations
Keywords
Metrics
Footnotes

**More Like This**

JSEG-based image segmentation in computer vision for agricultural mobile robot navigation
2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation - (CIRA)
Published: 2009

Visual content layer for scalable object recognition in urban image databases
2009 IEEE International Conference on Multimedia and Expo
Published: 2009

**Show More**

IEEE Personal Account
CHANGE USERNAME/PASSWORD

Purchase Details
PAYMENT OPTIONS
VIEW PURCHASED DOCUMENTS

Profile Information
COMMUNICATIONS PREFERENCES
PROFESSION AND EDUCATION
TECHNICAL INTERESTS

Need Help?
US & CANADA: +1 800 678 4333
WORLDWIDE: +1 732 981 0060
CONTACT & SUPPORT

Follow

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | IEEE Ethics Reporting | Sitemap | IEEE Privacy Policy

A not-for-profit organization, IEEE is the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity.

© Copyright 2024 IEEE - All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies.

**IEEE Account**

» Change Username/Password

» Update Address

**Purchase Details**

» Payment Options

» Order History

» View Purchased Documents

**Profile Information**

» Communications Preferences

» Profession and Education

» Technical Interests

**Need Help?**

» **US & Canada:** +1 800 678 4333

» **Worldwide:** +1 732 981 0060

» Contact & Support

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | Sitemap | Privacy & Opting Out of Cookies

A not-for-profit organization, IEEE is the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity.

© Copyright 2024 IEEE - All rights reserved. Use of this web site signifies your agreement to the terms and conditions.

**Appendix MERLER02**

# Recognizing Groceries *in situ* Using *in vitro* Training Data

Michele Merler
*Telecommunications Engineering*
*University of Trento, Italy*
*michele.merler@studenti.unitn.it*

Carolina Galleguillos
*Computer Science & Engineering*
*University of California, San Diego*
*cgallegu@cs.ucsd.edu*

Serge Belongie
*Computer Science & Engineering*
*University of California, San Diego*
*sjb@cs.ucsd.edu*

## Abstract

*The problem of using pictures of objects captured under ideal imaging conditions (here referred to as* in vitro*) to recognize objects in natural environments* (in situ) *is an emerging area of interest in computer vision and pattern recognition. Examples of tasks in this vein include assistive vision systems for the blind and object recognition for mobile robots; the proliferation of image databases on the web is bound to lead to more examples in the near future. Despite its importance, there is still a need for a freely available database to facilitate study of this kind of training/testing dichotomy. In this work one of our contributions is a new multimedia database of 120 grocery products, GroZi-120. For every product, two different recordings are available:* in vitro *images extracted from the web, and* in situ *images extracted from camcorder video collected inside a grocery store. As an additional contribution, we present the results of applying three commonly used object recognition/detection algorithms (color histogram matching, SIFT matching, and boosted Haar-like features) to the dataset. Finally, we analyze the successes and failures of these algorithms against product type and imaging conditions, both in terms of recognition rate and localization accuracy, in order to suggest ways forward for further research in this domain.*

## 1. Introduction

Object detection and recognition are important tasks in computer vision. Nearly all of the algorithms that address these problems demand large amounts of labeled training examples for each object. Moreover, it is generally necessary that the training examples be acquired under environmental conditions (illumination, pose, etc.) that are representative of what will be observed in the testing data.

In many real-world applications, however, having access to training data for which the appearance is drawn from the same distribution as that of the testing data is not practical. Examples of such applications include assistive vision technology for the blind and object recognition for mobile robots. Consider for example the Semantic Robot Vision Challenge (SRVC)[1], a new research competition designed to push the state-of-the-art in image understanding and automatic acquisition of knowledge from large unstructured databases of images (such as those generally found on the web). Central to the challenge is the problem of training a "laptop-on-wheels" object recognizer using only a textual list of objects (provided shortly before the competition) and access to an internet image search engine. In the case of our application of interest, we seek to design a portable vision system for helping blind people find groceries in a supermarket, with a starting point again consisting of a textual list – in this case a shopping list – and access to publicly available grocery images on the internet.

The web contains an immense collection of structured and unstructured image databases and offers the potential to generate useful models for image recognition and localization. In order to explore this potential in at least one application domain, we built a multimedia database consisting of 120 different grocery items easily retrievable from the web. For every product, two different recordings are available: *in vitro* images extracted from the web (captured under ideal imaging conditions), and *in situ* images extracted from camcorder video collected inside a grocery store. We refer to a grocery product with its corresponding UPC code as an object. Each product has several different image examples that were extracted from the web and from video captures. The intent of this dataset is to serve as a seed upon which the set of images can grow dynamically by user interaction in the future. In particular, methods such as [5] could be used to discover and label large amounts of *in situ* object examples in photos that currently lack textual annotation using *in vitro* training data from online resources such as Amazon Groceries.

Our goal in this paper is to provide a highly varied real world multimedia database for studying object recognition and detection with the above-described distinction between training and testing data, as well as baseline performance

[1] http://www.cs.cmu.edu/ ∼prybski/SRVC/

**1-4244-1180-7/07/$25.00 ©2007 IEEE**



Authorized licensed use limited to: Penn State University. Downloaded on June 13,2024 at 18:52:04 UTC from IEEE Xplore. Restrictions apply.

of several approaches used to solve object localization and recognition. In Section 2 we first review the state of the art object detection and recognition algorithms used in our baseline study, and then we discuss existing databases of common use in the computer vision community. In Section 3 we introduce the GroZi-120 database, which is publicly available, with a detailed description of what it contains and how it was created. Section 4 presents the features used and the different models for localization and recognition. Section 5 shows the results on the database and finally in Section 6 we discuss the results and propose ideas for future work.

## 2. Related work

### 2.1. Detection/recognition algorithms

Color histogram matching is one of the first algorithms ever applied to detect and recognize objects in images and videos. Swain and Ballard's early work on color object recognition by means of fast matching color histograms by intersecting them [20] opened the way to many different approaches with a common ground. Computational complexity, which has always been one major bottleneck of the histogram extraction and comparison based search tasks, has recently been overcome with the introduction of the integral histogram by Fatih Porikli [19]. On the other hand, the choice of which colorspace or chrominance plane to use is still an open issue [14].

In addition to color features, one can employ methods based on on shape [4] and/or gray-scale object descriptors. There are two distinct problems linked to such approaches: interest point detection and description of the distribution of smaller-scale features within the interest point neighborhood. The most commonly used interest point detector is probably the Harris corner detector [12], while for the second problem, David Lowe's SIFT [16] descriptor has been shown to outperform the others, as it is invariant to scale and rotation transformations. This property can be explained by the fact that it captures a substantial amount of information about the spatial intensity patterns, while at the same time being robust to small deformations or localization errors.

In the framework of object detection, a major contribution has been offered by Viola and Jones [21], who introduced a fast and reliable classifier. They suggested using a cascade of classifiers, each of which is a combination of a subset of weak learners selected via Adaboost [8]. The weak learners use Haar-like features, which are responses of filters computed extremely quickly by employing integral images.

### 2.2. Computer vision databases

Over the years many datasets have been introduced in the computer vision community, in order to provide a mean of evaluation for the different algorithms developed. Here we present a sample of recent ones, summarizing their characteristics. In this way we will make evident the innovative contribution offered by the GroZi-120 dataset, which we will introduce in Section 3.

The Pascal Object Recognition Database Collection [1] consists of data gathered for an object recognition in natural scenes challenge that took place in 2005 and 2006. In its final version, the collection contains 5,304 images, provided by Microsoft Research Cambridge and collected from the photo-sharing web-site Flickr, of 10 object classes. All images are annotated with instances of all the classes, for a total of 9,507 labelled objects. The set presents variability of scale, pose, background clutter and degree of occlusion for every object.

Caltech 101, now expanded into Caltech 256 [7], contains 30607 images, grouped into 256 object classes, with a mean of 119 images per class. It also includes a special category for clutter and background. It is widely used for object recognition but is not recommended for object localization.

Another example is SOIL-47 [13], a database of household objects, many of the same shape, viewed over a significant portion of the viewing sphere. The images show mainly multicolored objects, many of them consisting of planar surfaces (boxes) and with generally complex color structure. The database contains 24 objects with approximately planar surfaces and 22 complex scenes. Both objects and scenes are presented against a black background, in absence of clutter. Three different kinds of appearance variation are included: 3D viewpoint, illumination intensity and occlusion/distractors. A database like SOIL-47 may represent an extremely rich *in vitro* training set for the problem we investigate, but it lacks the *in situ* counterpart that is needed for testing.

A more recent computer vision database is the Amsterdam Library of Object Images (ALOI) [9], a collection of one-thousand small objects. The creators of the dataset systematically varied viewing angle, illumination angle, and illumination color for each object, and additionally captured wide-baseline stereo images. It includes over a hundred images per object, yielding a total of 110,250 images. Again, no clutter or complex background are present.

Finally, the ETH-80 [15] database consists of 80 objects from 8 chosen categories captured in high-resolution color images, with segmentation masks provided for every image. Each object is represented by 41 images from viewpoints spaced equally over the upper viewing hemisphere.

Apart from the Caltech 256 and the PASCAL dataset, all the reported databases present a uniform one-color background, so objects are easily segmentable from the background (which is usually black and presents no clutter). Moreover, in almost all cases no occlusion is present and


Authorized licensed use limited to: Penn State University. Downloaded on June 13,2024 at 18:52:04 UTC from IEEE Xplore. Restrictions apply.

the sizes and position of objects inside the images are normalized, as pointed out in [18]. In contrast, we propose a dataset that contains images that present a significant difference in quality (between training and testing) and where individual objects have both cluttered and uniform background. Therefore it can be used for object localization as well as recognition approaches, and it can be used for training and testing purposes, either with images or video captures.

## 3. GroZi-120 database

The GroZi-120[2] is a multimedia database of 120 grocery products. It was created in order to address the problem of learning using training data that differs in quality from the testing data for object recognition and localization tasks. The objects belonging to it vary in color, size, opacity, shape and rigidity, as can be seen in Figure 2. The dataset introduces variabilities not systematically offered in previous available collections. In fact, many of our testing images contain multiple instances of the same object, which may present partial occlusion and truncation, as well as size and orientation variations. Furthermore, the location of the product varies considerably from image to image and different objects are found in the same frames. All these properties are exemplified in Figure 4. Every product has two different representations in the database: one captured *in vitro* and another *in situ*. The *in vitro* images are isolated and captured under ideal imaging conditions (e.g.,stock photography studio or a lab) and they can be found on the web, more specifically in grocery web stores such as Froogle[3].

In order to obtain *in vitro* representations we queried the web using a list of approximately 4000 UPC codes associated with the products of a local grocery store. A Perl script was generated to automatically crawl the web looking for images of the products by using the UPC code. In case the system did not respond positively to a query, a textual description of the product was obtained from the UPC online database[4], and the query was performed again with the new keywords. A manual selection of the downloaded instances was executed in order to eliminate duplicates and images representing products different from the desired one. In order to make the *in vitro* images usable as a training set for any algorithm, we set their background to transparent and we obtained a binary mask in order to extract only the useful information out of every image, as shown in Figure 5 (a).

| | *in vitro* | *in situ* |
|---|---|---|
| Total number of images | 676 | 11194 |
| Average number of images per object | 5.6 | 93.3 |
| Min. number of images per object | 2 | 14 |
| Max. number of images per object | 14 | 814 |

Table 1. General statistics of *in situ* and *in vitro* images for the GroZi-120 database. The reduced number of *in vitro* samples arises from the difficulty to retrieve unique instances of a given product.



Figure 1. Sample of *in vitro* images of multiple views for a product.




Figure 2. Sample of *in vitro* images for different products.




Figure 3. Sample of *in situ* images for the above products.

Therefore the *in vitro* images become easy to analyze and at the same time, coming from different online vendors and stock photo suppliers, they include a variety of illuminations, sizes and poses. Figure 1 shows the different views of a particular product represented by *in vitro* images. In contrast, *in situ* representations come from natural environments (real world). Figure 3 shows the *in situ* representations of the objects in Figure 2. We went to the grocery store that provided us the UPC list (a campus convenience store), and shot 29 videos on the same day at 30 fps, encoded as Divx 5.2.1 with a bitrate of 2000 kbps using a VGA resolution MiniDV camcorder, for a total of around 30 minutes of footage. Such videos include every product present in the *in vitro* part of the dataset. These images were selected every

[2]http://grozi.calit2.net
[3]http://www.froogle.com
[4]http://www.upcdatabase.com


Authorized licensed use limited to: Penn State University. Downloaded on June 13,2024 at 18:52:04 UTC from IEEE Xplore. Restrictions apply.

5 frames and were stored together with their location in the video (video number, frame number, rectangle coordinates), as shown in Figure 5 (b). These images present variations in scale, illumination, reflectance, color, pose and rotation, while the video frames provide a cluttered background. One of the benefits of this data is that it represents the typical low quality of a real world image. Hence, different algorithms for object recognition (where *in vitro* images can be compared to *in situ* images) and object localization (search for products in videos) can be tested on this dataset. Table 1 reports statistics about the dataset. Grozi-120 constitutes a database of mixed multimedia content, with images and videos representing the same objects under different visual conditions, together with their text annotations. The creation of such a dataset involved different multimedia resources, spanning from the web to digital camcorders.

## 4. Object detection and recognition algorithms

In order to provide a baseline characterization of the level of difficulty inherent to this problem domain, we tested a selection of popular object detection and recognition approaches and studied their performance on the GroZi-120 dataset. Below we describe the different features used and the approaches implemented in detail.

### 4.1. Features and dissimilarity measures

**Color Histogram:** We first tested our database with chrominance planes belonging to 3 different color spaces: YCbCr, HSV and Lab. A preliminary study showed that the ab plane from Lab provided the best results. We computed histograms of 16 bins per channel, a and b, calculated separately, for a total of 32 bins. Then we generated a histogram of the a and b channels for every *in vitro* image in the dataset. Histograms belonging to the same product were subsequently averaged, bin per bin, in order to obtain a final template histogram representative of the object.

**SIFT:** We computed SIFT [16] keypoints for every *in vitro* image in the data set in order to represent images using scale and rotation invariant descriptors. The keypoints were computed using binaries provided by the UCLA Vision Lab[5]. Each grocery product was represented by a "bag of keypoints" extracted from the *in vitro* images corresponding to a particular object. Therefore the calculated keypoints correspond to the different views of the same object without including background. The background pixels were set to zero when the masks were applied to the images. The product shape information is implicitly (albeit weakly) captured by the descriptors, since they are computed on the masked images.

**Boosted Haar-like features:** We used the implementation in the Haar training utility of the Intel OpenCV library[6]. Positive samples were synthetically created from the *in vitro* images by applying randomly generated perspective distortions, until 200 positive samples were obtained, including the *in vitro* instances. The dimensions of such images were obtained by computing the average ratio of all the masked *in vitro* images and then resizing them to have a longest dimension of 50 pixels.




Figure 4. *In situ* video frame sample. There are 2 instances of product 103, product 4 truncated, 2 instances of object 33 (one almost completely occluded by the other), 1 sample of product 27 (rotated out of plane) and 1 instance of product 95.




Figure 5. (a) Binary mask applied to web samples. (b) Product image cropped from video frame and stored together with coordinates.

### 4.2. Recognition

In our recognition study, *in vitro* images were used as training data, while *in situ* images were used as test data. In color histogram matching we computed color histograms for every *in situ* image in the dataset. Then the distances between the *in vitro* image template and *in situ* image histograms were calculated, according to 3 different metrics: Euclidean, $\chi^2$ and histogram intersection ($L_1$ distance). Once the distances were computed, we calculated the ROC curves shown in Section 5 by integrating a bidimensional histogram of the distances. We obtained the best performance using color histogram intersection.

[5] http://vision.ucla.edu/ ~vedaldi/code/sift/sift.html

[6] http://www.intel.com/technology/computing/opencv/


Authorized licensed use limited to: Penn State University. Downloaded on June 13,2024 at 18:52:04 UTC from IEEE Xplore. Restrictions apply.

Recognition using SIFT proceeded similarly to the approach used for color histogram matching. For each product we computed the bag of features obtained from its *in vitro* images. Then we matched the features with the keypoints of every *in situ* image in the dataset. The distances between samples were represented by the number of matches between the *in vitro* and the *in situ* one. The ROC curves were computed in the same manner as for the color histogram matching.

With regard to training in Adaboost, the cascade was allowed to be 14 stages deep, with a maximum false alarms rate (FAR) of 0.5 per stage.The ROC curves were obtained by computing the distances to the *in situ* images resized to the dimensions of the training set of the particular product. Such distances were computed for the classifier by taking the weighted sums of the responses of the features selected at every stage and finding their difference from the thresholds of the stage. All the differences were then summed together to obtain a unique value for the distance.

### 4.3. Localization

Localization experiments were conducted by trying to identify the location of products present in video sequences, using *in vitro* images as the training set. Testing was performed for every product using 14 frames out of the videos containing the product, with its locations manually identified as ground truth, and 100 frames from the same videos not containing any of the objects in the dataset. We took into account the possible presence of multiple instances of the product in the same frame.

We used two different metrics to evaluate localization: a yes/no rule from which we obtain true positives and false positives rates, and a metric based on the average object area recall and the average detected area precision as defined in [17], that gives us the *overall recall* and *overall precision* rates. The first is a frame-based metric where a $yes$ is given if the center of the detected box (meaning the best match) lies within the ground truth region and a $no$ otherwise, as presented in [3]. In the second, the recall for an object is defined by the authors as the proportion of its area that is covered by the algorithm's output boxes for every frame and the overall recall is computed as the weighted average recall of all frames. Precision of an output box is defined as the proportion of its area that covers the ground truth objects and overall precision is the weighted average precision of all frames.

The color histogram matching approach relies on the integral histogram computation as in [19]. We compute the *in vitro* template histogram in the same way as in recognition. Then, we performed a frame by frame analysis first computing the integral histogram (for the a and b channels) of the whole frame and then moving a window at 5 different scales in raster scan order around the frame, computing the color histograms of different regions. Then those histograms are intersected with the *in vitro* template. If the distance is accepted by the system, it is stored to be later compared to all the other best matches in the frame. Finally a maximum of 6 windows are kept, which correspond to the best scores.

The SIFT approach for object localization consisted first in computing the bag of keypoints for every *in vitro* object and for each frame as in recognition. Since the *in vitro* image sizes are different, we normalized the coordinates of all the keypoints. Then, we matched the keypoints against the frame features as in [16]. In order to reduce the number of outliers and locate the object we centered a circle in the *in vitro* average image, with a diameter equal to the average image diagonal. Iterating over the matches we kept the circle containing the maximum number of matches. Subsequently we computed the centroid of the locations of the corresponding matching features in the frame, and also their average distance with respect to the centroid. Using the previous pair of matches, we found the ones inside a new circle with a diameter equal to the average distance in the *in vitro* and frame instances. If the number of frame matches found is greater or equal than the *in vitro* ones found we consider the object as detected, otherwise we use the matches that were not taken into account in the first place and proceed in the same way. If in both cases the condition is not satisfied, we do not consider the object as detected. This approach was performed for different circles so we could handle multiple instances of a product on a frame.

In the Adaboost based method, the classifier obtained from the set of boosted Haar-like features during the training phase is used to decide whether a series of rectangles analyzed at different scales within each frame contains the product of interest. This process is performed by comparing the responses of the filters selected for every stage of the classifier during the training phase to thresholds also selected during the training.

## 5. Experiments and Results

In this Section we present the results for recognition and localization of the algorithms presented in the previous sections. For recognition, we chose 10 *in vitro* image samples for every product for training, corresponding to the highest number of SIFT keypoints, considering that a higher number of keypoints often means an image with lower blurriness. Synthetic images were created when the original number of samples was less than 10. Subsequently we chose 10 *in situ* images per object for testing using the same criteria, but not considering synthetic images. In Figure 6 we report the average performances. It can be noticed how the overall best performing algorithm is SIFT, while the color histogram matching curve (CHM), after an initial flat behavior, rapidly grows towards a 99 percent match rate. Finally, Adaboost offers the worst performance. One of the reasons for


Authorized licensed use limited to: Penn State University. Downloaded on June 13,2024 at 18:52:04 UTC from IEEE Xplore. Restrictions apply.

such a difference relies on the lack of a consistent number of original samples for training, alleviated by the generation of synthetic images, which do not offer effectively diverse characteristics. SIFT and CHM, on the other hand, need only very few samples to extract the necessary information to perform recognition. Figures 7 and 8 present two interesting cases. In the case of product 1, there is a clear disparity between the performances of the different algorithms. While the SIFT performed well, probably due to the distinctiveness of the text and symbols on the product's box, the color histogram curve is affected by the predominance of the white color, which is not particularly distinguishable in the chrominance plane, and performs poorly since it is based on a more global approach. On the other hand, in Figure 8, CHM benefits from the multimodal histogram obtained from a product presenting colors (yellow, orange and green) very well distinguishable in the ab plane, while SIFT and Haar-like features encounter a lack of distinctive feature points on most of the product's surface.

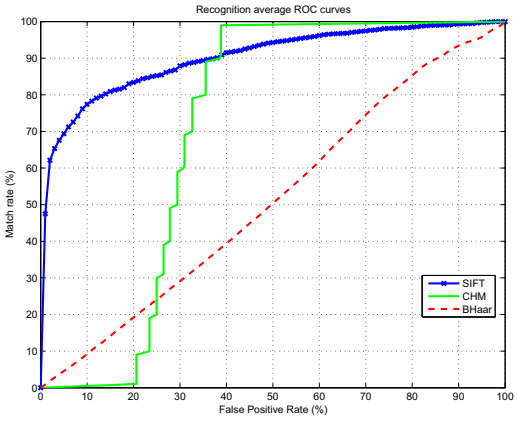



Figure 6. Average ROC curves for all the Products.

In the case of localization we used 100 frames with no instances of any product in the database to test the False Positive rates. Table 2 shows the localization rates in percentages for the three different algorithms. We use a yes/no rule to obtain true and false positives, based on whether the center of the detected box lies or not within ground truth region. Recall is defined as the proportion of the ground truth area covered by the detection boxes. Precision is defined as the proportion of the detection boxes that covers the ground truth area. The best and worst cases are selected considering the combination of all the 4 rates. Figure 9 presents the best and worst cases. Looking the results we can notice that SIFT outperforms the other methods, followed by CHM. For Adaboost the discussion reported for recognition still applies. The high values of the standard deviations in the table show that the methods do not perform uniformly over the set of products. This fact is due to the complexity and variety of the dataset, which presents challenges of variable difficulty for any algorithm. Furthermore from the experiments emerges a discrepancy between the efficacy of the three methods in recognition and localization. While recognition yields decent results, the average performances in localization are poor.

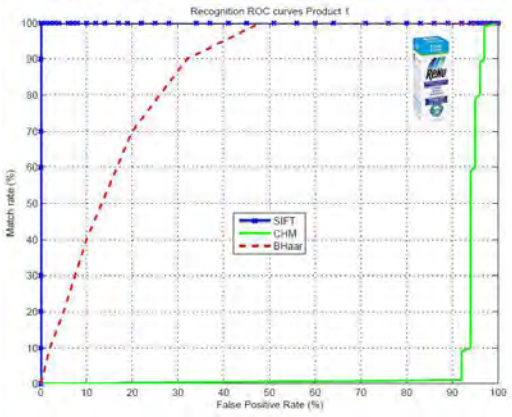



Figure 7. ROC curves for Product 1.

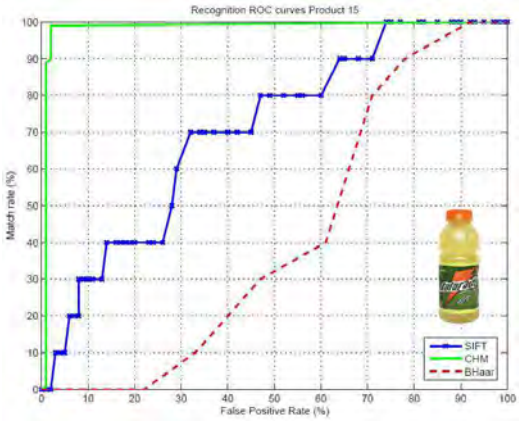



Figure 8. ROC curves for Product 15.

Such a discrepancy is due to the different nature of the tasks: in recognition we are still operating in a controlled environment, where although it is true that the samples we are comparing come from the different *in vitro* and *in situ* worlds, it is also true that those instances are segmented from the cluttered background, and therefore easier to analyze. On the other hand, localization must consider the full background, which contributes a significant amount of confusion to the problem. In this sense localization expresses the core of the problem: find the correspondence between *in vitro* samples and cluttered, noise corrupted, realistic *in situ* scenes. Figures 10, 11 and 12 are examples of different cases of the localization performance of the 3 algorithms on the dataset. From Figure 10 we see CHM perform well for product 52, while SIFT and boosted Haar-like features are not as good. In fact, object 52 has a distinctive orange color, which clearly stands out of the background. On the other hand, the remaining two methods cannot rely on such precious information and they are misled by the text on the


Authorized licensed use limited to: Penn State University. Downloaded on June 13,2024 at 18:52:04 UTC from IEEE Xplore. Restrictions apply.

neighboring products. Figure 11 shows good localization by all the methods on the product where SIFT performs best. The characteristics of the product (clear and distinctive colors, multiple pattern variations) generate informative features for every algorithm. Only boosted Haar like features are partially misled by a neighbor. Taking a closer look at the misdetection and a fraction of an *in vitro* sample, the similarity between the two is evident.

| **CHM** | % Recall | % Precision | % TP | % FP |
|---|---|---|---|---|
| Mean | 15 | 17 | 18 | 65 |
| Std Dev. | 28 | 16 | 35 | 32 |
| Best (20) | 71 | 82 | 100 | 4 |
| Worst(32) | 1.7 | 0.2 | 0 | 100 |

| **SIFT** | % Recall | % Precision | % TP | % FP |
|---|---|---|---|---|
| Mean | 72 | 18 | 22 | 62 |
| Std Dev. | 20 | 17 | 26 | 28 |
| Best (34) | 14 | 83 | 93 | 25 |
| Worst (9) | 26 | 0.9 | 0 | 64 |

| **ADA** | % Recall | % Precision | % TP | % FP |
|---|---|---|---|---|
| Mean | 15 | 17 | 18 | 65 |
| Std Dev. | 13 | 13 | 19 | 24 |
| Best (92) | 35 | 74 | 50 | 38 |
| Worst (5) | 0.5 | 0.2 | 0 | 92 |

Table 2. Localization performances of the 3 algorithms. Mean, Standard Deviation, Best Case (product number in brackets) and Worst Case percentages are reported for Overall Recall, Overall Precision, True Positives and False Positives.




Figure 9. Best and Worst localization samples for CHM, SIFT and ADA. The products 32, 9 and 5 correspond to the worst results and 20, 34 and 92 the best.

Finally we present a bad localization example: product 74. CHM is misled by portions of the frame very similar in color to some parts of the *in vitro* image. For SIFT, the image is too rich with keypoints for the algorithm to be able to succeed, in particular because the product of interest is mainly composed by uniform patterns, so that the keypoints are found on corners in the writing. This together with the reflections on the palstic surface of the Yoohoo bottle prove deceptive for the algorithm. Boosted Haar like features on the other hand, finds (in grayscale) a product which is dark on the top and the bottom, and light in the middle, with some texture inside.

## 6. Discussion and future work

Our contributions in this paper include (a) a new multimedia database – GroZi-120 – for studying object recognition *in situ* (i.e., sitting in the real world) using training images from an *in vitro* source (i.e., captured under ideal conditions) and (b) baseline performance figures of three widely used recognition/detection algorithms that highlight the challenge presented by this database. The Grozi-120 database contains both *in situ* and *in vitro* representations of the same products, and it presents a wide range of variability for each product with respect to size, color, rigidity, shape, illumination, viewpoint and quality.

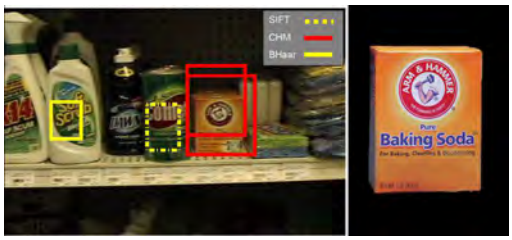



Figure 10. Localization example for Product 52. On the right: *in vitro* image sample.




Figure 11. Example of good localization performance by all the methods for Product 34. The *in vitro* image detail on the top right resembles the only false positive.

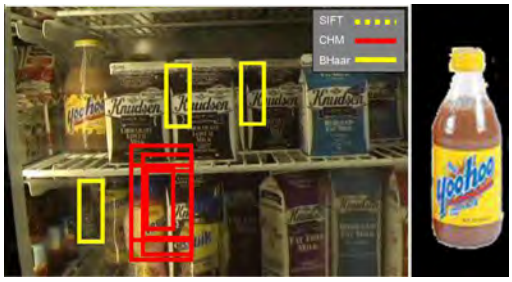



Figure 12. Example of poor localization performance for Product 74. On the right: *in vitro* image sample.

Gathering useful training information from images captured in ideal conditions is linked to the semantic web image retrieval issue, addressed among others by [22]. In this work the authors demonstrate the application of common image search metrics applied to images captured with a camera-equipped mobile device to find matching images on the Web or other general-purpose databases. The results of the testbed suggest the need for more precise and elaborate detection/recognition algorithms to solve the proposed problem. In fact, the aim of our work was to provide a baseline against which to test new methods. Color histogram matching may benefit from the works on color invariants, e.g., [11] and [10]. Another interesting direction may consist in a combination of the grayscale and color based methods, such as [2] and [6]. Our database offers not only a


Authorized licensed use limited to: Penn State University. Downloaded on June 13,2024 at 18:52:04 UTC from IEEE Xplore. Restrictions apply.

testbed for such approaches trying to link real world data to clean web or studio images, it can also be seen as an intermediate step or a bridge between the two representations of the same object. In fact, web and multimedia retrieval can benefit from a set of pre-existing labeled samples, while at the same time dynamically increasing and improving such a set. This type of problem can find applications in mobile robot navigation like the Semantic Robot Vision Challenge (SRVC) and also in assistive technologies for the visually impaired . In fact, in the case of a blind or visually impaired person that uses a device that recognizes products in a grocery store, it would be impractical to acquire *in situ* data every time we need to train the system, thus the *in vitro* data captured from the web is a good source of training data.

Therefore we intend as future work to use the *in vitro* data set as a seed to build upon, as the user base continues to use the database and expanding the *in situ* part as a means of evaluating different algorithms. Furthermore, we plan to include more objects in the database and expand the number examples per object. We also plan to fuse recognition and detection methods in order to improve the results as well as to make use of context information about physical object proximity, identifying products nearby on the shelf to improve localization of objects in natural scenes.

**Acknowledgments**

The authors would like to thank Vincent Rabaud, Stephan Steinbach, Anton Escobedo, Al Labotski, Robert Meza and Michael McQuen for valuable help and feedback, as well as all the people involved in the GroZi project at UCSD. Michele Merler was supported by the California Institute for Telecommunications and Information Technology (Calit2) 2006 Summer Undergraduate Research Scholarship Program. Partial support was also provided by NSF CAREER Grant #0448615 and the Alfred P. Sloan Research Fellowship.

## References

[1] The pascal object recognition database collection. 2005. 2

[2] A. E. Abdel-Hakim and A. A. Farag. CSIFT: A SIFT descriptor with color invariant characteristics. In *CVPR06*, pages 1978–1983, 2006. 7

[3] F. Bashir and F. Porikli. Performance evaluation of object detection and tracking systems. *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, June 2006. 5

[4] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, pages 509–522, 2002. 2

[5] N. Ben-Haim, B. Babenko, and S. Belongie. Improving web-based image search via content based clustering. *International Workshop on Semantic Learning Applications in Multimedia (SLAM), New York City.*, 2006. 1

[6] T. Burghardt and J. Calic. Analysing animal behaviour in wildlife videos using face detection and tracking. *IEEE Proceedings - Vision, Image, and Signal Processing*, 153(3):305–312, June 2006. 7

[7] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. *IEEE. CVPR 2004, Workshop on Generative-Model Based Vision*, 2004. 2

[8] Y. Freund and R. Schapire. Experiments with a new boosting algorithm. *Proceedings of the Thirteenth International Conference on Machine Learning, Morgan Kaufmann.*, page 148156, 1996. 2

[9] J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders. The Amsterdam library of object images. *Int. J. Comput. Vision*, 61(1):103–112, 2005. 2

[10] J.-M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and A. Dev. Color and scale: The spatial structure of color images. In *ECCV*, pages 331–341, 2000. 7

[11] T. Gevers and H. Stokman. Robust histogram construction from color invariants for object recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(1):113–117, 2004. 7

[12] C. Harris and M. Stephens. A combined edge and corner detector. *4th Alvey Vision Conference*, pages 189–192, 1988. 2

[13] D. Koubaroulis, J. Matas, and J. Kittler. Evaluating colour-based object recognition algorithms using the soil-47 database. *Asian Conference on Computer Vision, 2002.*, 2002. 2

[14] J. Y. Lee and S. I. Yoo. An elliptical boundary model for skin color detection. *Proc of the International Conference on Imaging Science, Systems, and Technology*, 2002. 2

[15] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. *IEEE CVPR'03*, June 2003. 2

[16] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60, 2*, pages 91–110, 2004. 2, 4, 5

[17] V. Y. Mariano, J. Min, J.-H. Park, R. Kasturi, D. Mihalcik, H. Li, D. Doermann, and T. Drayer. Performance evaluation of object detection algorithms. *ICPR*, 03:30965, 2002. 5

[18] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. Torralba, C. K. I. Williams, J. Zhang, and A. Zisserman. *Dataset Issues in Object Recognition*. Toward Category-Level Object Recognition, Springer-Verlag Lecture Notes in Computer Science., 2006. 3

[19] F. Porikli. Integral histogram: A fast way to extract histograms in cartesian spaces. *CVPR Vol. 1*, pages 829–836, 2005. 2, 5

[20] M. J. Swain and D. H. Ballard. Color indexing. *Int. J. Comput. Vision*, 7(1):11–32, 1991. 2

[21] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR (1)*, pages 511–518, 2001. 2

[22] T. Yeh, K. Tollmar, and T. Darrell. Searching the web with mobile images for location recognition. *CVPR*, 02:76–81, 2004. 7



Authorized licensed use limited to: Penn State University. Downloaded on June 13,2024 at 18:52:04 UTC from IEEE Xplore. Restrictions apply.

**Appendix MITCHELL01**

```
leader  01247cam a2200277 a 4500
001     9916982513406236
005     20040227073035.0
008     970211s1997    nyua     b    001 0 eng
010     ##$a   97007692
020     ##$a0070428077
035     ##$a(lcb)97007692
035     ##$9ARK2502C1
035     ##$a1698251
035     ##$a(ppiu)1698251-pittdb
035     ##$a(pittdb)1698251-pittdb
040     ##$aDLC $cDLC $dDLC $dOrLoB-B $dOCoLC
050     00$aQ325.5 $b.M58 1997
082     00$a006.3/1 $221
100     1#$aMitchell, Tom M. $q(Tom Michael), $d1951-
245     10$aMachine Learning / $cTom M. Mitchell.
260     ##$aNew York : $bMcGraw-Hill, $cc1997.
300     ##$axvii, 414 p. : $bill. ; $c25 cm.
504     ##$aIncludes bibliographical references and indexes.
505     00$g1. $tIntroduction --  $g2. $tConcept Learning and the General-to-Specific Ordering --
650     #0$aMachine learning.
650     #0$aComputer algorithms.
999     ##$aPWBO 05/14/98
948     ##$alcfc:nl;cm:mi
906     ##$aULS
```



BOOK

## Machine Learning

Mitchell, Tom M. (Tom Michael), 1951- / Tom M. Mitchell.
c1997
xvii, 414 p. : ill. ; 25 cm.

Available at Science and Engineering Library (Benedum Hall) General Collection 2 (Q325.5 .M58 1997) >

Report a Problem

TOP
SEND TO
GET IT
DETAILS
VIRTUAL BROW...

### Send to

CITATION EMAIL ENDNOTE/ZOTERI PRINT PERMALINK

### Get It

Please sign in to check if there are any request options. Sign in

BACK TO LOCATIONS

LOCATION ITEMS

Science and Engineering Library (Benedum Hall)
Available , General Collection 2 ; Q325.5 .M58 1997
(1 copy, 1 available, 0 requests)

Item in place
Loanable

### Details

| | |
|---|---|
| Title | Machine Learning |
| Creator | Mitchell, Tom M. (Tom Michael), 1951- > |
| Language | English |
| Publisher | New York : McGraw-Hill |
| Date | c1997 |
| Type | Book |
| Description | xvii, 414 p. : ill. ; 25 cm. |
| Notes | Includes bibliographical references and indexes. |
| Contents | 1. Introduction -- 2. Concept Learning and the General-to-Specific Ordering -- 3. Decision Tree Learning -- 4. Artificial Neural Networks -- 5. Evaluating Hypotheses -- 6. Bayesian Learning -- 7. Computational Learning Theory -- 8. Instance-Based Learning -- 9. Genetic Algorithms -- 10. Learning Sets of Rules -- 11. Analytical Learning -- 12. Combining Inductive and Analytical Learning -- 13. Reinforcement Learning. |
| Subject | Machine learning > Computer algorithms > |
| Identifier | LC : 97007692 ISBN : 0070428077 |
| Source | Library Catalog |

### Related reading

Recommended items that are related to the record



**Artificial intelligence : a** ...
Russell, Stuart J.
ISBN 9780131038059
suggested by



**The Elements of Statistical** ...
Hastie, Trevor.
ISBN 9780387848587
suggested by



**Programming collective** ...
Segaran, Toby.
ISBN 9780596550684
suggested by



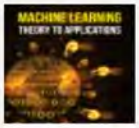
**Pattern recognition and** ...
Bishop, Christopher M.
ISBN 9780387310732
suggested by



**Introduction to algorithms**
Cormen, Thomas H.
ISBN 9780262033848
suggested by

More

## Virtual Browse



Machine learning : theory to applications ...
2022



Machine Learning ...
c1997



Machine Learning and Interpretation in Neuroimagin ...
2012



Modeling, machine learning and astronomy : First ...
2020



Moc App in M Lear Cogi
202(

## Links

Find this Item in Google Books ›
Display Source Record ›

## Syndetics Unbound

Summary | Librarian Lists | Reader Reviews
Tags | Browse Shelf

### SUMMARY

This book covers the field of machine learning, which is the study of algorithms that allow computer programs to automatically improve through experience. The book is intended to support upper level undergraduate and introductory level graduate courses in machine learning.

### LIBRARIAN RECOMMENDS

**Related Lists**




**AI - Artificial Intelligence**
by University for the Creative Arts




**Artificial Intelligence and Higher Education**
by Christian Schmidt, Special Projects Librarian

SEE MORE LISTS

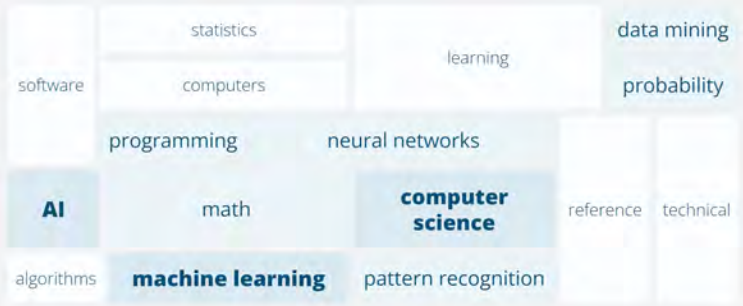
### READER REVIEWS

★★★☆☆ Reviewed by one person **What do you think?** Write your own review

This is an introductory book on Machine Learning. There is quite a lot of mathematics and statistics in the book, which I like. A large number of methods and algorithms are introduced: Neural Networks Bayesian Learning Genetic Algorithms Reinforcement Learning The material covered is very interesting and clearly explained. I find the presentation, however, a bit lacking. I think it has (read more)

Review by IvanIdris (LibraryThing) ★★★☆☆

TAGS




BROWSE TAGS

BROWSE SHELF




Q325.5 .M575 2011 Q325.5 .M63 2017

SEE FULL SHELF

*Catalog enrichment powered by Syndetics Unbound*

**Appendix MOVSHOVITZ01**

# MARC View

| | | | |
|---|---|---|---|
| LEADER 01154nam a22002652 4500 | | | |
| 001 20056824 | | | |
| 003 SIRSI | | | |
| 005 20170414130234.0 | | | |
| 006 m d | | | |
| 007 cr n | | | |
| 008 151014e20150607nju s\|\|\|\|\|\|\|\| 2\|eng\|d | | | |
| 020 | | | a\| 9781467369657 |
| 020 | | | a\| 1467369659 (Spiral) c\| USD 700.00 Retail Price (Publisher) 9\| Active Record |
| 037 | | | b\| 00066573 |
| 040 | | | a\| BIP US d\| WaSeSS |
| 110 | 2 | | a\| IEEE Staff e\| Author |
| 245 | 1 | 0 | a\| 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) h\| [electronic resource] |
| 260 | | | a\| Piscataway : b\| IEEE c\| June 2015 |
| 506 | | | a\| License restrictions may limit access. |
| 520 | 8 | | a\| Annotation b\| computer, vision, pattern, cvpr, machine, learning |
| 521 | | | a\| Scholarly & Professional b\| IEEE |
| 655 | | 0 | a\| Electronic books. |
| 856 | 4 | 0 | u\| http://SK8ES4MC2L.search.serialssolutions.com/?sid=sersol&SS_jc=TC0001767523&title=2015%20IEEE%20Conference%20on%20Computer%20Vision%20and%20Pattern%20Recognition%20%28CVPR%29 |
| 955 | | | a\| Books In Print, (c) 2017 R.R. Bowker LLC |
| 949 | | | a\| Electronic resource w\| ASIS c\| 1 i\| 20056824-1001 l\| ONLINE m\| ONLINE r\| Y s\| Y t\| ONLINE u\| 4/18/2017 o\| .STAFF. 360 MARC |

ASK A LIBRARIAN

PennState

Copyright @2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

Legal Statements | PSU Hotlines

PENN STATE UNIVERSITY LIBRARIES

Libraries Home (814) 865-6368
Libraries Intranet (Staff Only)
Accessibility Help
Website Feedback
Policies and Guidelines
Staff Directory

CONNECT WITH PENN STATE UNIVERSITY LIBRARIES

Facebook
Twitter
Instagram

# 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [electronic resource]

Corporate Author: IEEE Staff

Published: Piscataway : IEEE June 2015

Access Online

serialssolutions.com

ASK A LIBRARIAN

**Restrictions on Access:**
License restrictions may limit access.

**Summary:**
Annotation computer, vision, pattern, cvpr, machine, learning

**Genre(s):**
Electronic books

**ISBN:**
9781467369657
1467369659 (Spiral)

**Audience Notes:**
Scholarly & Professional IEEE

View MARC record | catkey: 20056824

PennState

Copyright @2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

Legal Statements | PSU Hotlines

PENN STATE UNIVERSITY LIBRARIES

Libraries Home (814) 865-6368
Libraries Intranet (Staff Only)
Accessibility Help
Website Feedback
Policies and Guidelines
Staff Directory

CONNECT WITH PENN STATE UNIVERSITY LIBRARIES

Facebook
Twitter
Instagram

IEEE.org IEEE *Xplore* IEEE SA IEEE Spectrum More Sites Donate Cart Create Account Personal Sign In

Browse My Settings Help

Access provided by: Penn State University Sign Out

Access provided by: Penn State University Sign Out

All

ADVANCED SEARCH

Conferences > 2015 IEEE Conference on Compu...

# Ontological supervision for fine grained classification of Street View storefronts

**Publisher: IEEE**

Cite This PDF

Yair Movshovitz-Attias ; Qian Yu ; Martin C. Stumpe ; Vinay Shet ; Sacha Arnoud ; Liron Yatziv All Authors

47 Cites in Papers

2 Cites in Patents

242 Full Text Views

Alerts

Manage Content Alerts

Add to Citation Alerts

Abstract

Document Sections

1. Introduction
2. Related Work
3. Challenges in Storefront Classification
4. Ontology Based Generation of Training Data
5. Model Architecture and Training

Show Full Outline

Authors

Figures

References

Citations

Keywords

Downl PDF

**Abstract:**
Modern search engines receive large numbers of business related, local aware queries. Such queries are best answered using accurate, up-to-date, business listings, that c... **View more**

**Metadata**

**Abstract:**
Modern search engines receive large numbers of business related, local aware queries. Such queries are best answered using accurate, up-to-date, business listings, that contain representations of business categories. Creating such listings is a challenging task as businesses often change hands or close down. For businesses with street side locations one can leverage the abundance of street level imagery, such as Google Street View, to automate the process. However, while data is abundant, labeled data is not; the limiting factor is creation of large scale labeled training data. In this work, we utilize an ontology of geographical concepts to automatically propagate business category information and create a large, multi label, training dataset for fine grained storefront classification. Our learner, which is based on the GoogLeNet/Inception Deep Convolutional Network architecture and classifies 208 categories, achieves human level accuracy.

**Published in:** 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

**Date of Conference:** 07-12 June 2015

**DOI:** 10.1109/CVPR.2015.7298778

**Date Added to IEEE *Xplore*:** 15 October 2015

**Publisher:** IEEE

Metrics

More Like This

ISBN Information:

ISSN Information:

Conference Location: Boston, MA, USA

Contents

## SECTION 1.
# Introduction

Following the popularity of smart mobile devices, search engine users today perform a variety of locality-aware queries, such as Japanese restaurant near me, Food nearby open now, or Asian stores in San Diego. With the help of local business listings, these queries can be answered in a way that is tailored to the user's location.

Creating accurate listings of local businesses is time consuming and expensive. To be useful for the search engine, the listing needs to be accurate, extensive, and importantly, contain a rich representation of the business category. Recognizing that a Japanese Restaurant is a type of Asian Store that sells Food, is essential in accurately answering a large variety of queries. Listing maintenance is a never ending task as businesses often move or close down. In fact it is estimated that 10% of establishments go out of business every year, and in some segments of the market, such as the restaurant industry, the rate is as high as 30% [24].

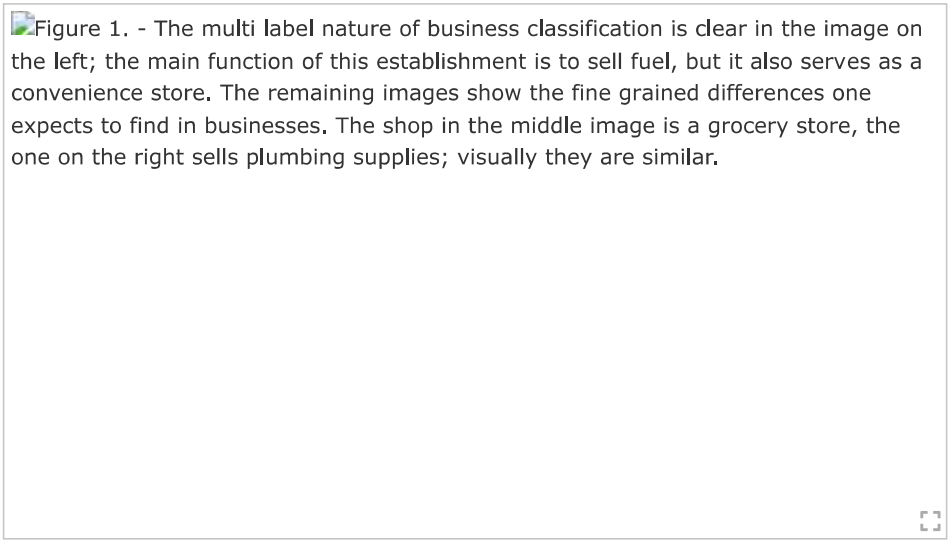



**Figure 1.**
The multi label nature of business classification is clear in the image on the left; the main function of this establishment is to sell fuel, but it also serves as a convenience store. The remaining images show the fine grained differences one expects to find in businesses. The shop in the middle image is a grocery store, the one on the right sells plumbing supplies; visually they are similar.

The turnover rate makes a compelling case for automating the creation of business listings. For businesses with a physical presence, such as restaurants and gas stations, it is a natural choice to use data from a collection of street level imagery. Probably the most recognizable such collection is Google Street View which contains hundreds of millions of 360° panoramic images, with geolocation information.

In this work we focus on business storefront classification from street level imagery. We view this task as a form of multi-label fine grained classification. Given an image of a storefront, extracted

from a Street View panorama, our system is tasked with providing the most relevant labels for that business from a large set of labels. To understand the importance of associating a business with multiple labels, consider the gas station shown in Figure 1 (left). While its main purpose is fueling vehicles, it also serves as a convenience or grocery store. Any listing that does not capture this subtlety will be of limited value to its users. Similarly, stores like Target or Walmart sell a wide variety of products from fruit to home furniture, all of which should be reflected in their listings. The problem is fine grained as business of different types can differ only slightly in their visual appearance. An example of such a subtle difference is shown in Figure 1. The middle image shows the front of a grocery store, while the image on the right is of a plumbing supply store. Visually they are similar. The discriminative information can be very subtle, and appear in varying locations and scales in the image; this, combined with the large number of categories needed to cover the space of businesses, require large amounts of training data.

The contribution of this work is two fold. First, we provide an analysis of challenges of a storefront classification system. We show that the intra-class variations can be larger than differences between classes (see Figure 2). Textual information in the image can assist the classification task, however, there are various drawbacks to text based models: Determining which text in the image belongs to the business is a hard task; Text can be in a language for which there is no trained model, or the language used can be different than what is expected based on the image location (see Figure 3). We discuss these challenges in detail in Section 3.

Finally, we propose a method for creating large scale labeled training data for fine grained storefront classification. We match street level imagery to known business information using both location and textual data extracted from images. We fuse information from an ontology of entities with geographical attributes to propagate category information such that each image is paired with multiple labels with different levels of granularity. Using this data we train a Deep Convolutional Network that achieves human level accuracy.

## SECTION 2.
# Related Work

The general literature on object classification is vast. Object category classification and detection [9] has been driven by the Pascal VOC object detection benchmark [8] and more recently the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [26]. Here, we focus on reviewing related work on analysis of street view data, fine-grained classification and the use of Deep Convolutional Networks.

### Analyzing Street View Data

Since its launch in 2007, Google Street View [28], [1] has been used by the computer vision community as both a test bed for algorithms [19], [31] and a source from which data is extracted and analyzed [12], [34], [21], [10], [6].

Early work on leveraging street level imagery focused on 3D reconstruction and city modeling. Cornelis et al. [6] focused on supplying textured 3D city models at ground level for car navigation system visualizations. Micusik et al. [21] used image segmentation cues and piecewise planar structures to build a robust 3D modeling system.

Later works have focused on extracting knowledge from Street View and leveraging it for particular tasks. In [34] the authors presented a system in which SIFT descriptors from 100,000 Street View images were used as reference data to be queried upon for image localization. Xiao et al. [31] proposed a multi view semantic segmentation algorithm that classified image pixels into high level categories such as ground, building, person, etc. Lee et al. [19] described a weakly supervised approach that mined midlevel visual elements and their connections in geographic data sets. Their approach discovered elements that vary smoothly over location. They evaluated their method using Street View images from the eastern coast of the United States. Their classifiers predicted location with a resolution of about 70 miles.

Most similar to our work, is that of Goodfellow et al. [12]. Both works utilize Street View as a map making source, and data mine information about real world objects. They focused on understanding street numbers, while we are concerned with local businesses. They described a method for street number transcription in Street View data. Their approach unified the localization, segmentation, and recognition steps by using a Deep Convolutional Network that operated directly on image pixels. The key idea behind their approach was to train a probabilistic model $P(S|X)$, where $S$ is a digit sequence, and $X$ an image patch, by maximizing $\log P(S|X)$ on a large training set. Their method, which was evaluated on tens of millions of annotated street number images from Street View, achieved above 90% accuracy and was comparable to human operators.

### Fine Grained Classification

Recently there has been renewed interest in Fine Grained classification [32], [33], [14] Yao et al. [33] modeled images by densely sampling rectangular image patches, and the interactions between pairs of patches, such as the intersection of the feature vectors of two image patches. In [32] the authors proposed a codebook-free representation which samples a large number of random patches from training images. They described an image by its response maps to matching the template patches. Branson et al. [4] and Wah et al. [29] proposed hybrid human-computer systems, which they described as a visual version of the *20-question game*. At each stage of the game, the algorithm chooses a question based on the content of the image, and previous user responses.

### Convolutional Networks

Convolutional Networks [11], [18] are neural networks that contain sets of nodes with tied parameters. Increases in size of available training data and availability of computational power, combined with algorithmic advances such as piecewise linear units [16], [13] and dropout training [15] have resulted in major improvements in many computer vision tasks. Krizhevsky et al. [17] showed a large improvement over state of the art in object recognition. This was later improved upon by Zeiler and Fergus [35], and Szegedy et al. [27].

On immense datasets, such as those available today for many tasks, overfitting is not a concern; increasing the size of the network provides gains in testing accuracy. Optimal use of computing resources becomes a limiting factor. To this end Dean et al. developed DistBelief [7], a distributed, scalable implementation of Deep Neural Networks. We base our system on this infrastructure.

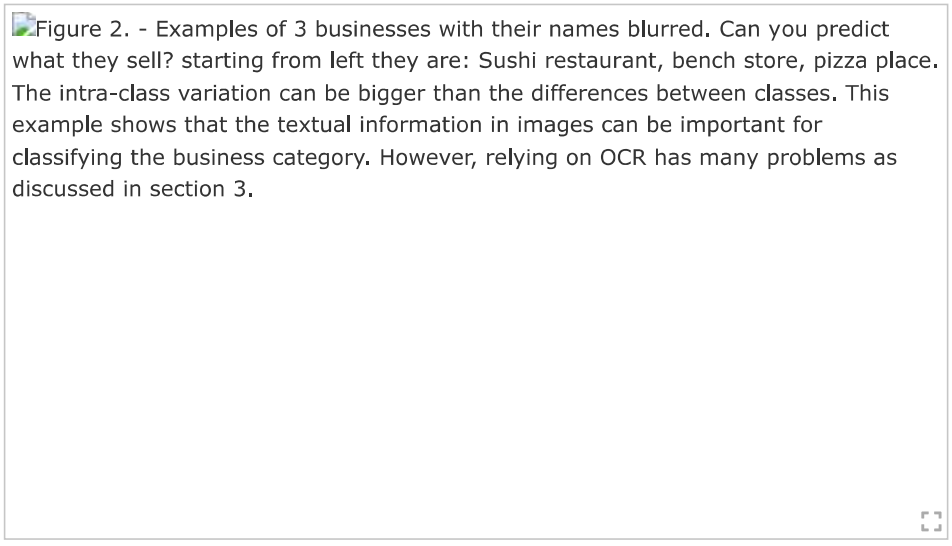



**Figure 2.**
Examples of 3 businesses with their names blurred. Can you predict what they sell? starting from left they are: Sushi restaurant, bench store, pizza place. The intra-class variation can be bigger than the differences between classes. This example shows that the textual information in images can be important for classifying the business category. However, relying on OCR has many problems as discussed in section 3.

## SECTION 3.
# Challenges in Storefront Classification

### Large within-Class Variance

Predicting the function of businesses is a hard task. The number of possible categories is large, and the similarity between different classes can be smaller than within class variability. Figure 2 shows three business storefronts. Their names have been blurred. Can you tell the type of the business without reading its name? Two of them are restaurants of some type, the third sells furniture, in particular store benches (middle image). It is clear that the text in the image can be extreme 1 y useful for the classification task in these cases.

### Extracted Text is Often Misleading

The accuracy of text detection and transcription in real world images has increased significantly over the last few years [30], [22], but relying on the ability to transcribe text has drawbacks. We would like a method that can scale up to be used on images captured across many countries and languages. When using extracted text, we need to train a dedicated model per language, this requires a lot of effort in curating training data. Operators need to mark the location, language and transcription of text in images. When using the system it would fail if a business had a different language than what we expect for its location or if we are missing a model for that language ( Figure 3a). Text can be absent from the image, and if present can be irrelevant to the type of the business. Relying on text can be misleading even when the language model is perfect; the text can come from a neighboring business, a billboard, or a passing bus ( Figure 3b). Lastly, panorama stitching errors may distort the text in the image and confuse the transcription process ( Figure 3c).

However, it is clear that the textual parts of the image do contain information that can be helpful. Ideally we would want a system that has all the advantages of using text information, without the drawbacks mentioned. In Section 6.3 we show that our system implicitly learns to use textual cues, but is more robust to these errors.

### Business Category Distribution

The natural distribution of businesses in the world exhibits a “long tail”. Some businesses (e.g. McDonalds) are very frequent, but most of the mass of the distribution is in the large number of businesses that only have one location. The same phenomena is also true of categories. Some labels have an order of magnitude more samples than others. For example, for the Food And Drink category which contains restaurants, bars, cafes, etc, we have 300,000 images, while for Laundry Ser-vice our data contains only 13,000 images. We note that a large part of the distribution's mass is in smaller categories.

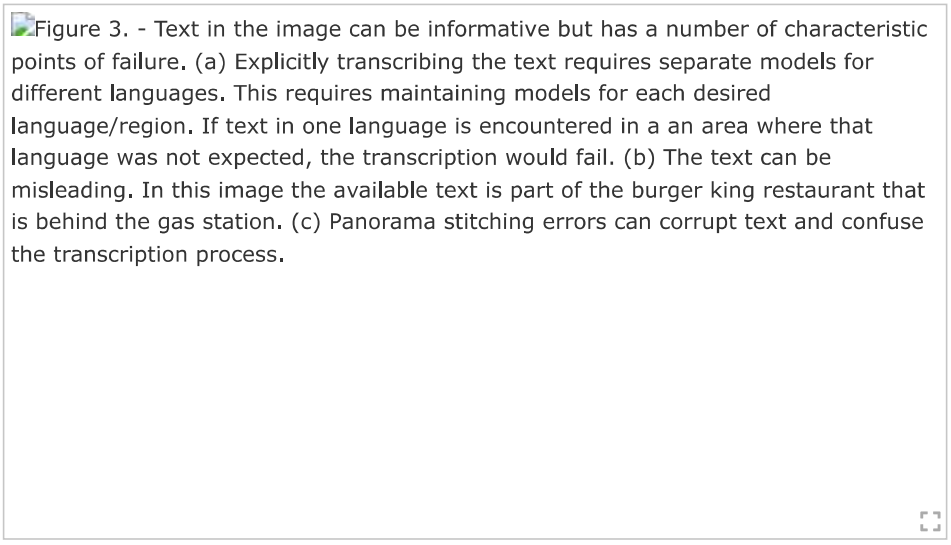



**Figure 3.**
Text in the image can be informative but has a number of characteristic points of failure. (a) Explicitly transcribing the text requires separate models for different

languages. This requires maintaining models for each desired language/region. If text in one language is encountered in a an area where that language was not expected, the transcription would fail. (b) The text can be misleading. In this image the available text is part of the burger king restaurant that is behind the gas station. (c) Panorama stitching errors can corrupt text and confuse the transcription process.

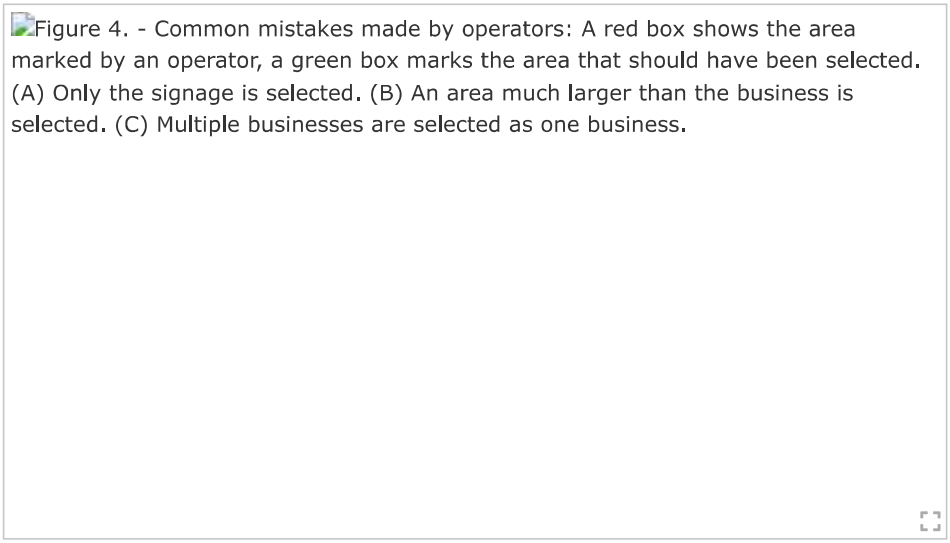



**Figure 4.**
Common mistakes made by operators: A red box shows the area marked by an operator, a green box marks the area that should have been selected. (A) Only the signage is selected. (B) An area much larger than the business is selected. (C) Multiple businesses are selected as one business.

### Labeled Data Acquisition

Acquiring a large set of high quality labeled data for training is a hard task in and of itself. We provide operators with Street View panoramas captured at urban areas in many cities across Europe, Australia, and the Americas. The operators are asked to mark image areas that contain business related information. We call these areas *biz-patches*. This process is not without errors. Figure 4 shows a number of common mistakes made by operators. The operators might mark only the business signage (4a), an area that is too large and contains unneeded regions (4b), multiple businesses in the same biz-patch (4c).

## SECTION 4.
# Ontology Based Generation of Training Data

Learning algorithms require training data. Deep Learning methods in particular are known for their need of large quantities of training instances, without which they overfit. In this section we describe a process for collecting a large scale training set, coupled with ontology-based labels.

Building a training set requires matching extracted biz-patches $p$ and sets of relevant category labels. First, we match a biz-patch with a particular business instance from a database of previously known businesses $\mathcal{B}$ that was manually verified by operators. We use the textual information and geographical location of the image to match it to a business. We detect text areas in the image, and transcribe them using an OCR software. This process suffers from the drawbacks of extracting text, but is useful for creating a set of candidate matches. This provides us with a set $\mathcal{S}$ of text strings. The biz-patch is geolocated and we combine the location information with the textual data. For each known business $b \in \mathcal{B}$, we create the same description, by combining its location and the set $\mathcal{T}$ of all the textual information that is available for it; name, phone number, operating hours, etc.

We decide that $p$ is a biz-patch of $b$ if geographical distance between them is less than approximately one city block, and enough extracted text from $\mathcal{S}$ matches $\mathcal{T}$

Using this technique we create a set of 3 million pairs $(p,\ b)$. However, due to the factors that motivated our work, the quality and completeness of the information varies greatly between businesses. For many businesses we do not have category information. Moreover, the operators who created the database were inconsistent in the way they selected categories. For example, a McDonalds can be labeled as a Hamburger Restaurant, a Fast Food Restau-rant, a Take Away Restaurant, etc. It is also plausible to label it simply as Restaurant. Labeling similar businesses with varying labels will confuse the learner.

We address this in two ways. First, by defining our task as a multi label problem we teach the classifier that many categories are plausible for a business. This, however, does not fully resolve the issue - When a label is missing from an example, the image is effectively used as *a negative* training instance for that label. It is important that training data uses a consistent set of labels for similar businesses. Here we use a key insight: the different labels used to describe a business represent different levels of specificity. For example, a hamburger restaurant is *a* restaurant. There is a containment relationship between these categories. On-tologies are a commonly used resource, holding hierarchical representations of such containment relations [3], [23]. We use an ontology that describes containment relationships between entities with a geographical presence, such as Restaurant, Pharmacy, and Gas Station. Our ontology, which is based on Google Map Maker's ontology, contains over 2,000 categories. For a pair $(p,\ b)$ for which we know the category label $c$, we locate $c$ in the ontology. We follow the containment relations described by the ontology, and add higher-level categories to the label set of $p$. The most general categories we consider are: Entertain-ment & Recreation, Health & Beauty, Lodging, Nightlife, Professional Services, Food & Drink, Shopping. Figure 5 shows an illustration of this process on a snippet from the ontology. Starting from an Italian Restaurant, we follow containment relations up predecessors in the ontology, until Food & Drink is reached.

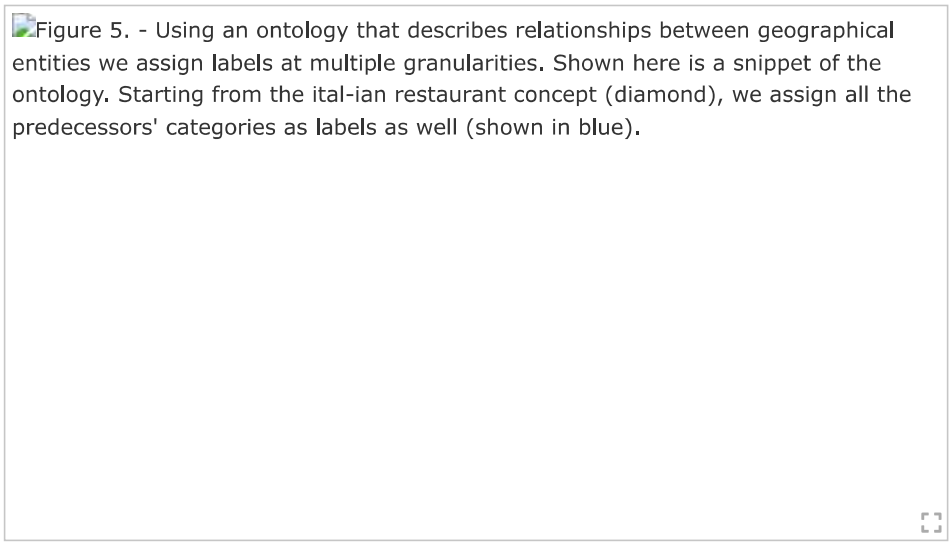



**Figure 5.**
Using an ontology that describes relationships between geographical entities we assign labels at multiple granularities. Shown here is a snippet of the ontology. Starting from the ital-ian restaurant concept (diamond), we assign all the predecessors' categories as labels as well (shown in blue).

This creates a large set of pairs $(p,\ s)$ where $p$ is a biz-patch and $s$ is a matching set of labels with varying levels of granularity. To ensure there is sufficient training data per label we omit labels whose frequency is very low and are left with 1.3 million biz-patches and 208 unique labels.

## SECTION 5.
# Model Architecture and Training

We base our model architecture on the winning submission for the ILSVRC 2014 classification and detection challenges by Szegedy et al. named GoogLeNet [27]. The model expands on the Network-in-Network idea of Lin et al. [20] while incorporating ideas from the theoretical work of Arora et al. [2]. Szegedy et al. forgo the use of fully connected layers at the top of the network and, by forcing the network to go through dimensionality reduction in middle layers, they are able to design a model that is much deeper than previous methods, while dramatically reducing the number of learned parameters. We employ the DistBe-lief [7] implementation of deep neural networks to train the model in a distributed fashion.

We create a train/test split for our data such that 1.2 million images are used for training the network and the remaining 100,000 images are used for testing. As a business can be imaged multiple times from different angles, the splitting is location aware. We utilize the fact that Street View panoramas are geotagged. We cover the globe with two types of tiles. Big tiles with an area of 18 kilometers, and smaller ones with area of 2 kilometers. The tiling alternates between the two types of tiles, with a boundary area of 100 meters between adjacent tiles. Panoramas that fall inside a big tile are assigned to the training set, and those that are located in the smaller tiles are assigned to the test set. This ensures that businesses in the test set were never observed in the training set while making sure that training and test sets were sampled from the same regions. This splitting procedure is fast and stable over time. When new data is available and a new split is made, train/test contamination is not an issue as the geographical locations are fixed. This allows for incremental improvements of the system over time.

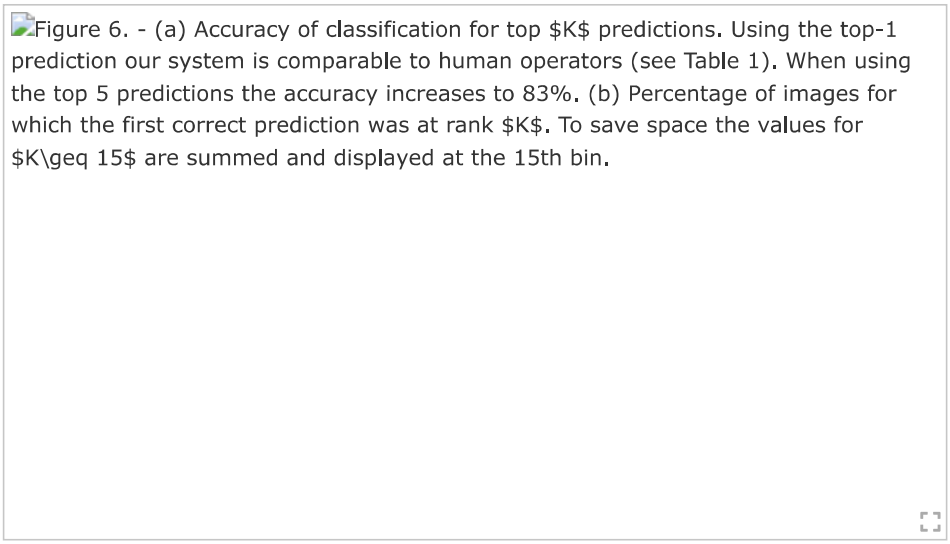



**Figure 6.**
(a) Accuracy of classification for top $K$ predictions. Using the top-1 prediction our system is comparable to human operators (see Table 1). When using the top 5 predictions the accuracy increases to 83%. (b) Percentage of images for which the first correct prediction was at rank $K$. To save space the values for $K \geq 15$ are summed and displayed at the 15th bin.

We first pre-train the network using images and ground truth labels from the ImageNet large scale visual recognition challenge with a Soft Max top layer, and once the model has converged we replace the top layer, and continue the training process with our business image data. This pretraining procedure has been shown to be a powerful initialization for image classification tasks [25], [5]. Each image is resized to $256 \times 256$ pixels. During training random crops of size $220 \times 220$ are given to the model as training images. We normalize the intensity of the images, add random photometric changes and create mirrored versions of the images to increase the amount of training data and guide the model to generalize. During testing a central box of size $220 \times 220$

pixels is used as input to the model. We set the network to have a dropout rate of 70% (each neuron has a 70% chance of not being used) during training, and use a Logistic Regression top layer. Each image is associated with all the labels found by the method described in Section 4. This setup is designed to push the network to share features between classes that are on the same path up the ontology.

## SECTION 6.
# Evaluation

In this section we describe our experimental results. We begin by providing a quantitative analysis of the system's performance, then describe two large scale human performance studies that show our system is competitive with the accuracy of human operators and conclude with quantitative results that provide understanding as to what features the system managed to learn.

When building a business listing it is important to have very high accuracy. If a listing contains wrong information it will frustrate its users. The requirements on coverage however can be less strict. If the category for some business images can not be identified, the decision can be postponed to a later date; each street address may have been imaged many times, and it is possible that the category could be determined from a different image of the business. Similarly to the work of Goodfellow et al. [12] on street number transcription, we propose to evaluate this task based on recall at certain levels of accuracy rather than evaluating the accuracy over all predictions. For automatically building listings we are mainly concerned with recall at 90% precision or higher. This allows us to build the listing incrementally, as more data becomes available, while keeping the overall accuracy of the listing high.

### 6.1. Fine Grained Classification Results

As described in section 4 each image is associated with one or more labels. We first evaluate the classifier's ability to retrieve at least one of those labels. For an image $i$, we define the ground truth label set $g_1$. The predictions $p_i$ are sorted by the classifier's confidence, and we define the *top-k* prediction set $p_i^k$ as the first $k$ elements in the sorted prediction list. A prediction for image $i$ is considered correct if $g_i \cap p_i^k \neq \emptyset$. Figure 6a shows the prediction accuracy as a function of labels predicted. The accuracy at *top-k* is shown for $k \in \{1, 3, 5, 7, 10\}$. Top-l performance is comparable to human annotators (see Section 6.2), and when the top 5 labels are used the accuracy increases to 83%. Figure 6b shows the distribution of first-correct-prediction, i.e. how far down the sorted list of predictions does one need to search before finding the first label that appears in $g_i$. We see that the first predicted label is by far the most likely and that the probability of having a predicted set $p_i^k$ that does not contain any of the true labels decreases with $k$. In order to save space we sum up all the probabilities for $k \in [15, 208]$ in one bin.

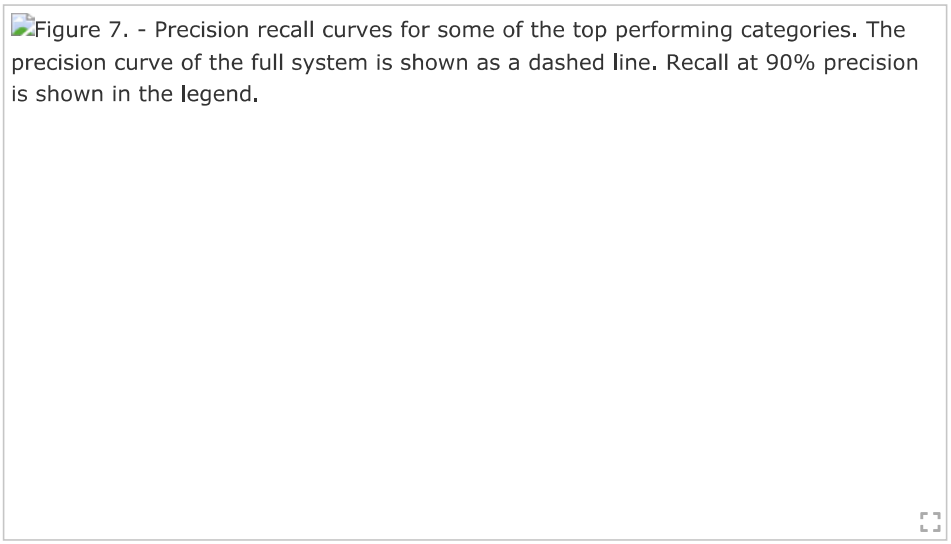

Figure 7. - Precision recall curves for some of the top performing categories. The precision curve of the full system is shown as a dashed line. Recall at 90% precision is shown in the legend.

**Figure 7.**
Precision recall curves for some of the top performing categories. The precision curve of the full system is shown as a dashed line. Recall at 90% precision is shown in the legend.

As mentioned above, one important metric for evaluation is recall at specific operating points. Figure 7 shows precision recall curves for some of the top performing categories and summary curve of the full system (dashed). The precision and recall of the full system is calculated by using the top-l prediction. For many categories we are able to recover the majority of businesses while precision is held at 90%. For a classification system to be useful in a practical setting the classifier's returned confidence must be well correlated with the quality of its prediction. Figure 8 shows a histogram of the number of correctly predicted labels in the top 5 predictions on a set of images whose labels were manually verified. The mean prediction confidence is indicated by color intensity (darker means higher confidence). Note the strong correlation between confidence and accuracy; for confidence above 80% normally at least 4 of the top labels are correct.

The GoogLeNet network [27] incorporates a number of new design elements that make it particularly appealing: by not using fully connected upper layers, and forcing the network to go through dimensionality reduction stages, the network has far fewer parameters while being much deeper than previous methods. We evaluate the use of this architecture by comparing it to the AlexNet network proposed by Krizhevsky et al. [17] on a 13 category, *single label* classification task. We select a set of useful categories that a user might search for, and train both networks to predict one label per image. Figure 9 shows recall at 0.9 precision on all categories for which at least one method has recall $\geq 0.1$. GoogLeNet outperforms the baseline for all categories, and for some, e.g. Pharmacy, recall more than doubles.

Figure 10 shows 30 sample images from the test set with their top 5 predictions. The model is able to classify these images with their high level categories and with fine grained, specialized classes. For example, the top-right image was classified by the model as: Beauty, Beauty Sa-lon, Cosmetics, Health Salon, and Nail Salon.

Figure 8. - Histogram of correct labels in the top 5 predictions for a set of 300 manually verified images. Color indicates mean prediction confidence. Note that the confidence in prediction is strongly correlated with the accuracy.

**Figure 8.**
Histogram of correct labels in the top 5 predictions for a set of 300 manually verified images. Color indicates mean prediction confidence. Note that the confidence in prediction is strongly correlated with the accuracy.

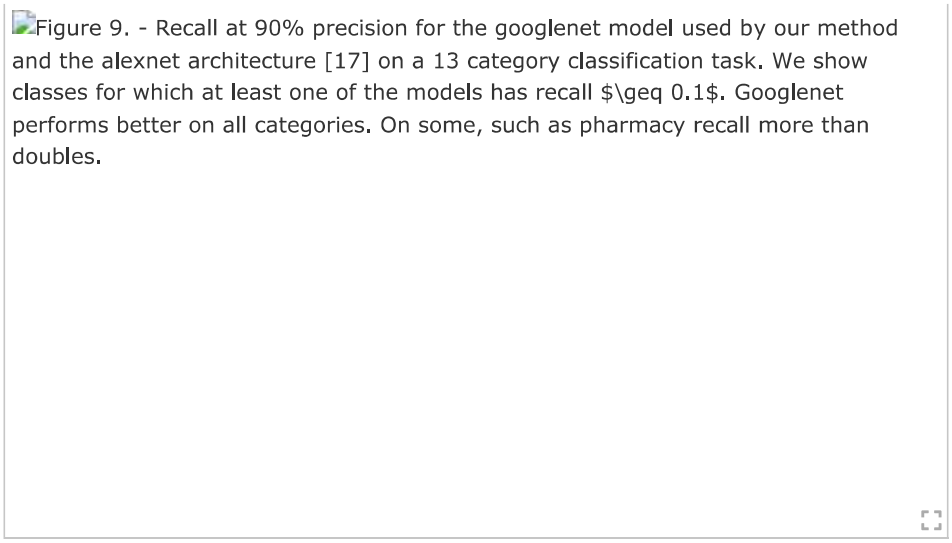
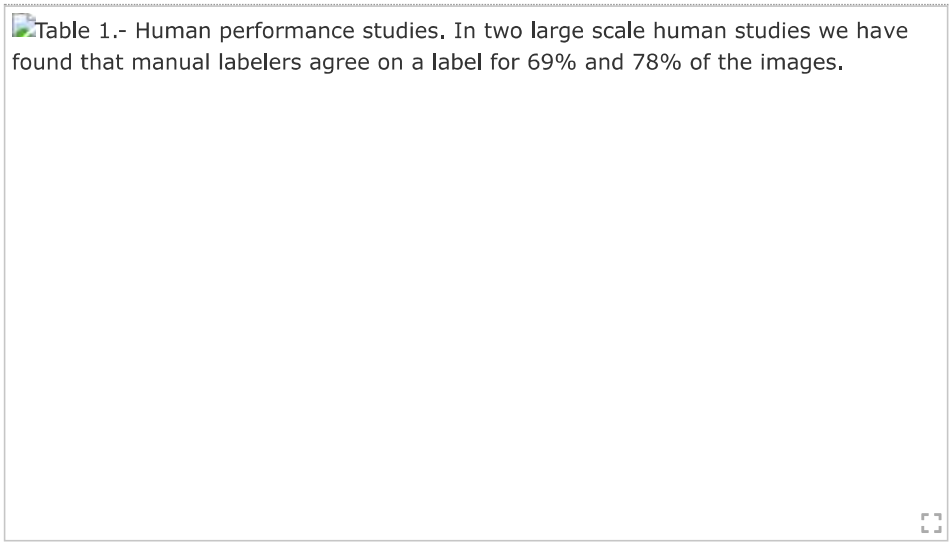



**Figure 9.**
Recall at 90% precision for the googlenet model used by our method and the alexnet architecture [17] on a 13 category classification task. We show classes for which at least one of the models has recall $\geq 0.1$. Googlenet performs better on all categories. On some, such as pharmacy recall more than doubles.

### 6.2. Human Performance Studies

Our system needs to perform at human level accuracy. Otherwise it will require a human verification post process. To estimate human accuracy on this task we have conducted two large scale human performance studies that check the agreement of operator-provided labels for the same business. In our experiments, subjects were shown images of businesses and selected one of 13 options (12 categories, and an Other category that stands for "none of the above".) Categories were chosen based on their financial impact for a business listing creator. The categories are: Food Store, *Restaurant/caffe*, Fi-nance, Pharmacy, Real Estate, Gas Station, Au-TOMOTIVE (other than gas stations), Fashion, Lodging, Laundry, Place Of Worship, Beauty. Note that the studies used full resolution images as opposed to the $256 \times 256$ images given to the algorithm. The first study had 73,272 images, each was shown to two operators. The operators agreed on 69% of image labels. In our second study we had 20,000 images, but each image was shown to *three to four* subjects. We found that the average agreement of the human operators was 78%. Table 1 shows a detailed summary of the human study results.

**Table 1.** Human performance studies. In two large scale human studies we have found that manual labelers agree on a label for 69% and 78% of the images.




### 6.3. Analysis: Learning to Read with Weak Labels

For some of the images in Figure 10, such as the image of the dental center (top row, second image from right) it is surprising that the model was capable of classifying it correctly. It is hard to think of a “canonical” dental center look, but even if we could, it doesn't seem likely that this image would be it. In fact, without reading the text, it seems impossible to correctly classify it. This suggests that the system has learned to use text when needed. Figure 11 shows images from Figure 10 for which we have manually blurred the discriminative text. Note especially the image of the dental center, and of the auto dealer. After blurring the test “Dental” the system is confused about the dental center; it believes it is a beauty salon of some sorts. However, for the auto dealer, it is still confident that this is a place that sells things, and is related to transportation and automotive. To take this experiment to its logical extreme, we also show a synthetic image, which contains only the word *Pharmacy*. The classifier predicts the relevant labels for it.

To us this is a compelling demonstration of the power of Deep Convolutional Networks to learn the correct representation for a task. Similar to a human in a country in which she does not know the language, it has done the best it can - learn that some words are correlated with specific types of businesses. Note that it was never provided with annotated text or language models. It was only provided with what we would consider as *weak textual labels*, images that contain text and labeled with category labels. Furthermore, when text is not available the system is able to make an accurate prediction if there is distinctive visual information.

## SECTION 7.
# Discussion

Business category classification, is an important part of location aware search. In this paper we have proposed a method for fine grained, multi label, classification of business storefronts from street level imagery. We show that our system learned to extract and associate text patterns in multiple languages to specific business categories without access to explicit text transcriptions. Moreover, our system is robust to the absence of text, and when distinctive visual information is available, it is able to make correct predictions. We show our system achieves human level accuracy.

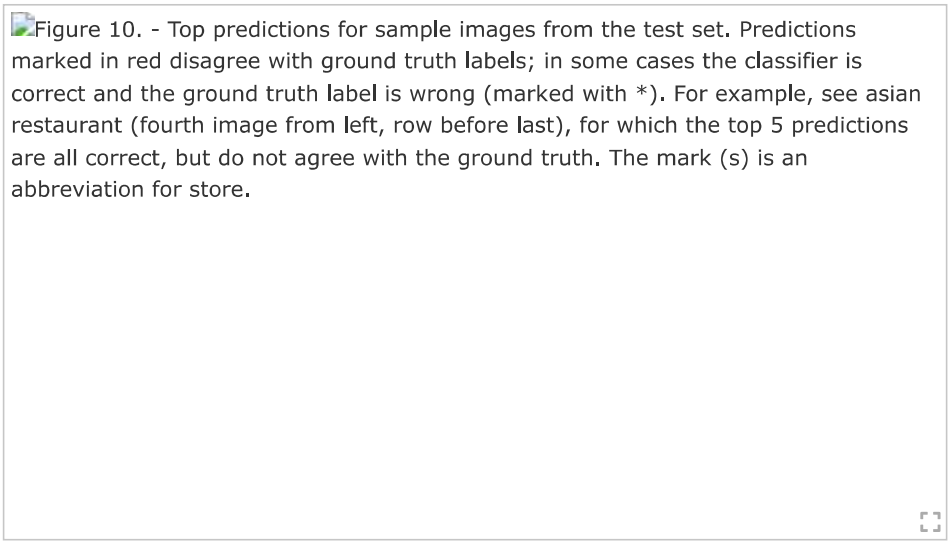



**Figure 10.**
Top predictions for sample images from the test set. Predictions marked in red disagree with ground truth labels; in some cases the classifier is correct and the ground truth label is wrong (marked with *). For example, see asian restaurant (fourth image from left, row before last), for which the top 5 predictions are all correct, but do not agree with the ground truth. The mark (s) is an abbreviation for store.

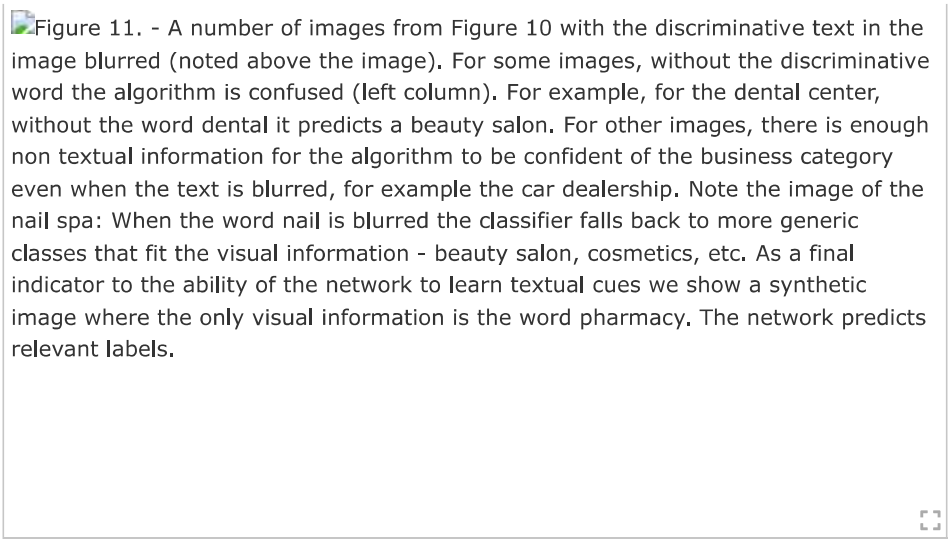



**Figure 11.**
A number of images from Figure 10 with the discriminative text in the image blurred (noted above the image). For some images, without the discriminative word the algorithm is confused (left column). For example, for the dental center, without the word dental it predicts a beauty salon. For other images, there is enough non textual information for the algorithm to be confident of the business category even when the text is blurred, for example the car dealership. Note the image of the nail spa: When the word nail is blurred the classifier falls back to more generic classes that fit the visual information - beauty salon, cosmetics, etc. As a final indicator to the ability of the network to learn textual cues we show a synthetic image where the only visual information is the word pharmacy. The network predicts relevant labels.

Using an ontology of entities with geographical attributes, we propagate label information during training, and produce a large set of 1.3 million images for a fine grained, multi label task. The use of non visual information, such as an ontology, to “ground” image data to real world entities is an exciting research direction, and there is much that can be done. For example, propagating information using the ontology at test time can increase both accuracy and recall. Node similarity in the ontology can be used to guide feature sharing between classes, and improve performance for seldom viewed classes.

## ACKNOWLEDGMENTS

The authors wish to thank Ian Goodfellow, Christian Szegedy, and Yaser Sheikh for helpful discussions.

Authors
Figures
References
Citations
Keywords
Metrics

**More Like This**

Three-Dimensional Displays: A Review and Applications Analysis
IEEE Transactions on Broadcasting
Published: 2011

Horizontal Resolution Enhancement of Autostereoscopy Three-Dimensional Displayed Image by Chroma Subpixel Downsampling
Journal of Display Technology
Published: 2015

**Show More**

IEEE Personal Account
CHANGE USERNAME/PASSWORD

Purchase Details
PAYMENT OPTIONS
VIEW PURCHASED DOCUMENTS

Profile Information
COMMUNICATIONS PREFERENCES
PROFESSION AND EDUCATION
TECHNICAL INTERESTS

Need Help?
US & CANADA: +1 800 678 4333
WORLDWIDE: +1 732 981 0060
CONTACT & SUPPORT

Follow

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | IEEE Ethics Reporting | Sitemap | IEEE Privacy Policy

A not-for-profit organization, IEEE is the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity.

© Copyright 2024 IEEE - All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies.

**IEEE Account**

» Change Username/Password

» Update Address

**Purchase Details**

» Payment Options

» Order History

» View Purchased Documents

**Profile Information**

» Communications Preferences

» Profession and Education

» Technical Interests

**Need Help?**

» **US & Canada:** +1 800 678 4333

» **Worldwide:** +1 732 981 0060

» Contact & Support

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | Sitemap | Privacy & Opting Out of Cookies

A not-for-profit organization, IEEE is the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity.

© Copyright 2024 IEEE - All rights reserved. Use of this web site signifies your agreement to the terms and conditions.

**Appendix MOVSHOVITZ02**

# Ontological Supervision for Fine Grained Classification of Street View Storefronts

Yair Movshovitz-Attias*, Qian Yu†, Martin C. Stumpe†, Vinay Shet†, Sacha Arnoud†, Liron Yatziv†

*Carnegie Mellon University
yair@cs.cmu.edu

†Google
{qyu,mstumpe,vinayshet,sacha,lirony}@google.com

## Abstract

*Modern search engines receive large numbers of business related, local aware queries. Such queries are best answered using accurate, up-to-date, business listings, that contain representations of business categories. Creating such listings is a challenging task as businesses often change hands or close down. For businesses with street side locations one can leverage the abundance of street level imagery, such as Google Street View, to automate the process. However, while data is abundant, labeled data is not; the limiting factor is creation of large scale labeled training data. In this work, we utilize an ontology of geographical concepts to automatically propagate business category information and create a large, multi label, training dataset for fine grained storefront classification. Our learner, which is based on the GoogLeNet/Inception Deep Convolutional Network architecture and classifies 208 categories, achieves human level accuracy.*

## 1. Introduction

Following the popularity of smart mobile devices, search engine users today perform a variety of locality-aware queries, such as *Japanese restaurant near me*, *Food nearby open now*, or *Asian stores in San Diego*. With the help of local business listings, these queries can be answered in a way that is tailored to the user's location.

Creating accurate listings of local businesses is time consuming and expensive. To be useful for the search engine, the listing needs to be accurate, extensive, and importantly, contain a rich representation of the business category. Recognizing that a JAPANESE RESTAURANT is a type of ASIAN STORE that sells FOOD, is essential in accurately answering a large variety of queries. Listing maintenance is a never ending task as businesses often move or close down. In fact it is estimated that 10% of establishments go out of business every year, and in some segments of the market, such as the restaurant industry, the rate is as high as 30% [24].




Figure 1. The multi label nature of business classification is clear in the image on the left; the main function of this establishment is to sell fuel, but it also serves as a convenience store. The remaining images show the fine grained differences one expects to find in businesses. The shop in the middle image is a grocery store, the one on the right sells plumbing supplies; visually they are similar.

The turnover rate makes a compelling case for automating the creation of business listings. For businesses with a physical presence, such as restaurants and gas stations, it is a natural choice to use data from a collection of street level imagery. Probably the most recognizable such collection is Google Street View which contains hundreds of millions of 360° panoramic images, with geolocation information.

In this work we focus on business storefront classification from street level imagery. We view this task as a form of multi-label fine grained classification. Given an image of a storefront, extracted from a Street View panorama, our system is tasked with providing the most relevant labels for that business from a large set of labels. To understand the importance of associating a business with multiple labels, consider the gas station shown in Figure 1 (left). While its main purpose is fueling vehicles, it also serves as a convenience or grocery store. Any listing that does not capture this subtlety will be of limited value to its users. Similarly, stores like Target or Walmart sell a wide variety of products from fruit to home furniture, all of which should be reflected in their listings. The problem is fine grained as business of different types can differ only slightly in their visual appearance. An example of such a subtle difference is shown in Figure 1. The middle image shows the front of a grocery store, while the image on the right is of a plumbing supply store. Visually they are similar. The discriminative infor-

978-1-4673-6964-0/15/$31.00 ©2015 IEEE

Authorized licensed use limited to: Penn State University. Downloaded on June 13,2024 at 18:51:02 UTC from IEEE Xplore. Restrictions apply.

mation can be very subtle, and appear in varying locations and scales in the image; this, combined with the large number of categories needed to cover the space of businesses, require large amounts of training data.

The contribution of this work is two fold. First, we provide an analysis of challenges of a storefront classification system. We show that the intra-class variations can be larger than differences between classes (see Figure 2). Textual information in the image can assist the classification task, however, there are various drawbacks to text based models: Determining which text in the image belongs to the business is a hard task; Text can be in a language for which there is no trained model, or the language used can be different than what is expected based on the image location (see Figure 3). We discuss these challenges in detail in Section 3.

Finally, we propose a method for creating large scale labeled training data for fine grained storefront classification. We match street level imagery to known business information using both location and textual data extracted from images. We fuse information from an ontology of entities with geographical attributes to propagate category information such that each image is paired with multiple labels with different levels of granularity. Using this data we train a Deep Convolutional Network that achieves human level accuracy.

## 2. Related Work

The general literature on object classification is vast. Object category classification and detection [9] has been driven by the Pascal VOC object detection benchmark [8] and more recently the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [26]. Here, we focus on reviewing related work on analysis of street view data, fine-grained classification and the use of Deep Convolutional Networks.

**Analyzing Street View Data**. Since its launch in 2007, Google Street View [28, 1] has been used by the computer vision community as both a test bed for algorithms [19, 31] and a source from which data is extracted and analyzed [12, 34, 21, 10, 6].

Early work on leveraging street level imagery focused on 3D reconstruction and city modeling. Cornelis et al. [6] focused on supplying textured 3D city models at ground level for car navigation system visualizations. Micusik et al. [21] used image segmentation cues and piecewise planar structures to build a robust 3D modeling system.

Later works have focused on extracting knowledge from Street View and leveraging it for particular tasks. In [34] the authors presented a system in which SIFT descriptors from $100,000$ Street View images were used as reference data to be queried upon for image localization. Xiao et al. [31] proposed a multi view semantic segmentation algorithm that classified image pixels into high level categories such as ground, building, person, etc. Lee et al. [19] described a weakly supervised approach that mined midlevel visual elements and their connections in geographic data sets. Their approach discovered elements that vary smoothly over location. They evaluated their method using Street View images from the eastern coast of the United States. Their classifiers predicted location with a resolution of about 70 miles.

Most similar to our work, is that of Goodfellow et al. [12]. Both works utilize Street View as a map making source, and data mine information about real world objects. They focused on understanding street numbers, while we are concerned with local businesses. They described a method for street number transcription in Street View data. Their approach unified the localization, segmentation, and recognition steps by using a Deep Convolutional Network that operated directly on image pixels. The key idea behind their approach was to train a probabilistic model $P(S|X)$, where $S$ is a digit sequence, and $X$ an image patch, by maximizing $\log P(S|X)$ on a large training set. Their method, which was evaluated on tens of millions of annotated street number images from Street View, achieved above 90% accuracy and was comparable to human operators.

**Fine Grained Classification**. Recently there has been renewed interest in Fine Grained classification [32, 33, 14] Yao et al. [33] modeled images by densely sampling rectangular image patches, and the interactions between pairs of patches, such as the intersection of the feature vectors of two image patches. In [32] the authors proposed a codebook-free representation which samples a large number of random patches from training images. They described an image by its response maps to matching the template patches. Branson et al. [4] and Wah et al. [29] proposed hybrid human-computer systems, which they described as a visual version of the *20-question game*. At each stage of the game, the algorithm chooses a question based on the content of the image, and previous user responses.

**Convolutional Networks**. Convolutional Networks [11, 18] are neural networks that contain sets of nodes with tied parameters. Increases in size of available training data and availability of computational power, combined with algorithmic advances such as piecewise linear units [16, 13] and dropout training [15] have resulted in major improvements in many computer vision tasks. Krizhevsky et al. [17] showed a large improvement over state of the art in object recognition. This was later improved upon by Zeiler and Fergus [35], and Szegedy et al. [27].

On immense datasets, such as those available today for many tasks, overfitting is not a concern; increasing the size of the network provides gains in testing accuracy. Optimal use of computing resources becomes a limiting factor. To this end Dean et al. developed DistBelief [7], a distributed, scalable implementation of Deep Neural Networks. We base our system on this infrastructure.



Authorized licensed use limited to: Penn State University. Downloaded on June 13,2024 at 18:51:02 UTC from IEEE Xplore. Restrictions apply.



Figure 2. Examples of 3 businesses with their names blurred. Can you predict what they sell? Starting from left they are: Sushi Restaurant, Bench store, Pizza place. The intra-class variation can be bigger than the differences between classes. This example shows that the textual information in images can be important for classifying the business category. However, relying on OCR has many problems as discussed in Section 3.

## 3. Challenges in Storefront Classification

**Large Within-Class Variance**. Predicting the function of businesses is a hard task. The number of possible categories is large, and the similarity between different classes can be smaller than within class variability. Figure 2 shows three business storefronts. Their names have been blurred. Can you tell the type of the business without reading its name? Two of them are restaurants of some type, the third sells furniture, in particular store benches (middle image). It is clear that the text in the image can be extremely useful for the classification task in these cases.

**Extracted Text is Often Misleading**. The accuracy of text detection and transcription in real world images has increased significantly over the last few years [30, 22], but relying on the ability to transcribe text has drawbacks. We would like a method that can scale up to be used on images captured across many countries and languages. When using extracted text, we need to train a dedicated model per language, this requires a lot of effort in curating training data. Operators need to mark the location, language and transcription of text in images. When using the system it would fail if a business had a different language than what we expect for its location or if we are missing a model for that language (Figure 3a). Text can be absent from the image, and if present can be irrelevant to the type of the business. Relying on text can be misleading even when the language model is perfect; the text can come from a neighboring business, a billboard, or a passing bus (Figure 3b). Lastly, panorama stitching errors may distort the text in the image and confuse the transcription process (Figure 3c).

However, it is clear that the textual parts of the image do contain information that can be helpful. Ideally we would want a system that has all the advantages of using text information, without the drawbacks mentioned. In Section 6.3 we show that our system implicitly learns to use textual cues, but is more robust to these errors.

**Business Category Distribution**. The natural distribution of businesses in the world exhibits a "long tail". Some businesses (e.g. McDonalds) are very frequent, but most of the mass of the distribution is in the large number of businesses that only have one location. The same phenomena is also true of categories. Some labels have an order of magnitude more samples than others. For example, for the FOOD AND DRINK category which contains restaurants, bars, cafes, etc, we have 300,000 images, while for LAUNDRY SERVICE our data contains only 13,000 images. We note that a large part of the distribution's mass is in smaller categories.



(a) Unexpected Language (b) Misleading Text (c) Stitching Errors

Figure 3. Text in the image can be informative but has a number of characteristic points of failure. (a) Explicitly transcribing the text requires separate models for different languages. This requires maintaining models for each desired language/region. If text in one language is encountered in a an area where that language was not expected, the transcription would fail. (b) The text can be misleading. In this image the available text is part of the Burger King restaurant that is behind the gas station. (c) Panorama stitching errors can corrupt text and confuse the transcription process.




(a) Area Too Small (b) Area Too Large (c) Multiple Businesses

Figure 4. Common mistakes made by operators: a red box shows the area marked by an operator, a green box marks the area that should have been selected. (A) Only the signage is selected. (B) An area much larger than the business is selected. (C) Multiple businesses are selected as one business.

**Labeled Data acquisition**. Acquiring a large set of high quality labeled data for training is a hard task in and of itself. We provide operators with Street View panoramas captured at urban areas in many cities across Europe, Australia, and the Americas. The operators are asked to mark image areas that contain business related information. We call these areas *biz-patches*. This process is not without errors. Figure 4 shows a number of common mistakes made by operators. The operators might mark only the business signage (4a), an area that is too large and contains unneeded regions (4b), multiple businesses in the same biz- patch (4c).



Authorized licensed use limited to: Penn State University. Downloaded on June 13,2024 at 18:51:02 UTC from IEEE Xplore. Restrictions apply.

## 4. Ontology Based Generation of Training Data

Learning algorithms require training data. Deep Learning methods in particular are known for their need of large quantities of training instances, without which they overfit. In this section we describe a process for collecting a large scale training set, coupled with ontology-based labels.

Building a training set requires matching extracted biz-patches $p$ and sets of relevant category labels. First, we match a biz-patch with a particular business instance from a database of previously known businesses $\mathcal{B}$ that was manually verified by operators. We use the textual information and geographical location of the image to match it to a business. We detect text areas in the image, and transcribe them using an OCR software. This process suffers from the drawbacks of extracting text, but is useful for creating a set of candidate matches. This provides us with a set $\mathcal{S}$ of text strings. The biz-patch is geolocated and we combine the location information with the textual data. For each known business $b \in \mathcal{B}$, we create the same description, by combining its location and the set $\mathcal{T}$ of all the textual information that is available for it; name, phone number, operating hours, etc. We decide that $p$ is a biz-patch of $b$ if geographical distance between them is less than approximately one city block, and enough extracted text from $\mathcal{S}$ matches $\mathcal{T}$.

Using this technique we create a set of 3 million pairs $(p, b)$. However, due to the factors that motivated our work, the quality and completeness of the information varies greatly between businesses. For many businesses we do not have category information. Moreover, the operators who created the database were inconsistent in the way they selected categories. For example, a McDonalds can be labeled as a HAMBURGER RESTAURANT, a FAST FOOD RESTAURANT, a TAKE AWAY RESTAURANT, etc. It is also plausible to label it simply as RESTAURANT. Labeling similar businesses with varying labels will confuse the learner.

We address this in two ways. First, by defining our task as a multi label problem we teach the classifier that many categories are plausible for a business. This, however, does not fully resolve the issue – When a label is missing from an example, the image is effectively used as a *negative* training instance for that label. It is important that training data uses a consistent set of labels for similar businesses. Here we use a key insight: the different labels used to describe a business represent different levels of specificity. For example, a hamburger restaurant *is a* restaurant. There is a containment relationship between these categories. Ontologies are a commonly used resource, holding hierarchical representations of such containment relations [3, 23]. We use an ontology that describes containment relationships between entities with a geographical presence, such as RESTAURANT, PHARMACY, and GAS STATION. Our ontology, which is based on Google Map Maker's ontology, contains over 2,000 categories. For a pair $(p, b)$ for which we know the category label $c$, we locate $c$ in the ontology. We follow the containment relations described by the ontology, and add higher-level categories to the label set of $p$. The most general categories we consider are: ENTERTAINMENT & RECREATION, HEALTH & BEAUTY, LODGING, NIGHTLIFE, PROFESSIONAL SERVICES, FOOD & DRINK, SHOPPING. Figure 5 shows an illustration of this process on a snippet from the ontology. Starting from an ITALIAN RESTAURANT, we follow containment relations up predecessors in the ontology, until FOOD & DRINK is reached.




Figure 5. Using an ontology that describes relationships between geographical entities we assign labels at multiple granularities. Shown here is a snippet of the ontology. Starting from the ITALIAN RESTAURANT concept (diamond), we assign all the predecessors' categories as labels as well (shown in blue).

This creates a large set of pairs $(p, s)$ where $p$ is a biz-patch and $s$ is a matching set of labels with varying levels of granularity. To ensure there is sufficient training data per label we omit labels whose frequency is very low and are left with 1.3 million biz-patches and 208 unique labels.

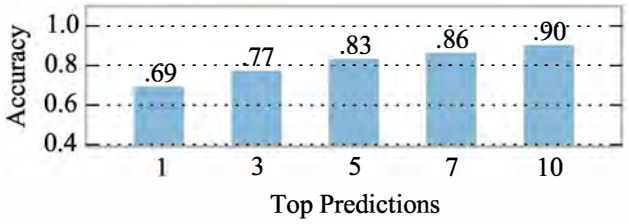
## 5. Model Architecture and Training

We base our model architecture on the winning submission for the ILSVRC 2014 classification and detection challenges by Szegedy et al. named GoogLeNet [27]. The model expands on the Network-in-Network idea of Lin et al. [20] while incorporating ideas from the theoretical work of Arora et al. [2]. Szegedy et al. forgo the use of fully connected layers at the top of the network and, by forcing the network to go through dimensionality reduction in middle layers, they are able to design a model that is much deeper than previous methods, while dramatically reducing the number of learned parameters. We employ the DistBelief [7] implementation of deep neural networks to train the model in a distributed fashion.

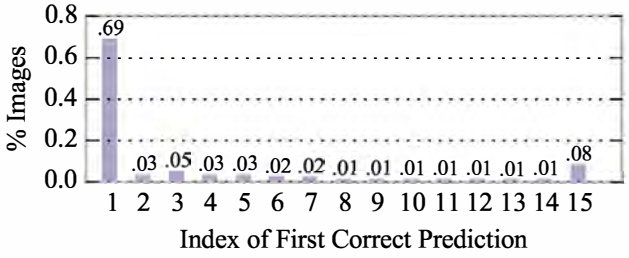
We create a train/test split for our data such that 1.2 million images are used for training the network and the remaining 100,000 images are used for testing. As a busi-



Authorized licensed use limited to: Penn State University. Downloaded on June 13,2024 at 18:51:02 UTC from IEEE Xplore. Restrictions apply.




(a) Accuracy at $K$




(b) First Correct Prediction

Figure 6. (a) Accuracy of classification for top $K$ predictions. Using the top-1 prediction our system is comparable to human operators (see Table 1). When using the top 5 predictions the accuracy increases to 83%. (b) Percentage of images for which the first correct prediction was at rank $K$. To save space the values for $K \geq 15$ are summed and displayed at the 15th bin.

ness can be imaged multiple times from different angles, the splitting is location aware. We utilize the fact that Street View panoramas are geotagged. We cover the globe with two types of tiles. Big tiles with an area of 18 kilometers, and smaller ones with area of 2 kilometers. The tiling alternates between the two types of tiles, with a boundary area of 100 meters between adjacent tiles. Panoramas that fall inside a big tile are assigned to the training set, and those that are located in the smaller tiles are assigned to the test set. This ensures that businesses in the test set were never observed in the training set while making sure that training and test sets were sampled from the same regions. This splitting procedure is fast and stable over time. When new data is available and a new split is made, train/test contamination is not an issue as the geographical locations are fixed. This allows for incremental improvements of the system over time.

We first pre-train the network using images and ground truth labels from the ImageNet large scale visual recognition challenge with a Soft Max top layer, and once the model has converged we replace the top layer, and continue the training process with our business image data. This pre-training procedure has been shown to be a powerful initialization for image classification tasks [25, 5]. Each image is resized to $256 \times 256$ pixels. During training random crops of size $220 \times 220$ are given to the model as training images. We normalize the intensity of the images, add random photometric changes and create mirrored versions of the images to increase the amount of training data and guide the model to generalize. During testing a central box of size $220 \times 220$ pixels is used as input to the model. We set the network to have a dropout rate of 70% (each neuron has a 70% chance of not being used) during training, and use a Logistic Regression top layer. Each image is associated with all the labels found by the method described in Section 4. This setup is designed to push the network to share features between classes that are on the same path up the ontology.

## 6. Evaluation

In this section we describe our experimental results. We begin by providing a quantitative analysis of the system's performance, then describe two large scale human performance studies that show our system is competitive with the accuracy of human operators and conclude with quantitative results that provide understanding as to what features the system managed to learn.

When building a business listing it is important to have very high accuracy. If a listing contains wrong information it will frustrate its users. The requirements on coverage however can be less strict. If the category for some business images can not be identified, the decision can be postponed to a later date; each street address may have been imaged many times, and it is possible that the category could be determined from a different image of the business. Similarly to the work of Goodfellow et al. [12] on street number transcription, we propose to evaluate this task based on recall at certain levels of accuracy rather than evaluating the accuracy over all predictions. For automatically building listings we are mainly concerned with recall at 90% precision or higher. This allows us to build the listing incrementally, as more data becomes available, while keeping the overall accuracy of the listing high.

### 6.1. Fine Grained Classification Results

As described in section 4 each image is associated with one or more labels. We first evaluate the classifier's ability to retrieve at least one of those labels. For an image $i$, we define the ground truth label set $g_i$. The predictions $p_i$ are sorted by the classifier's confidence, and we define the top-$k$ prediction set $p_i^k$ as the first $k$ elements in the sorted prediction list. A prediction for image $i$ is considered correct if $g_i \cap p_i^k \neq \emptyset$. Figure 6a shows the prediction accuracy as a function of labels predicted. The accuracy at top-$k$ is shown for $k \in \{1, 3, 5, 7, 10\}$. Top-1 performance is comparable to human annotators (see Section 6.2), and when the top 5 labels are used the accuracy increases to 83%. Figure 6b shows the distribution of first-correct-prediction, i.e. how far down the sorted list of predictions does one need to search before finding the first label that appears in $g_i$. We see that the first predicted label is by far the most likely and that the probability of having a predicted set $p_i^k$ that does not



Authorized licensed use limited to: Penn State University. Downloaded on June 13,2024 at 18:51:02 UTC from IEEE Xplore. Restrictions apply.

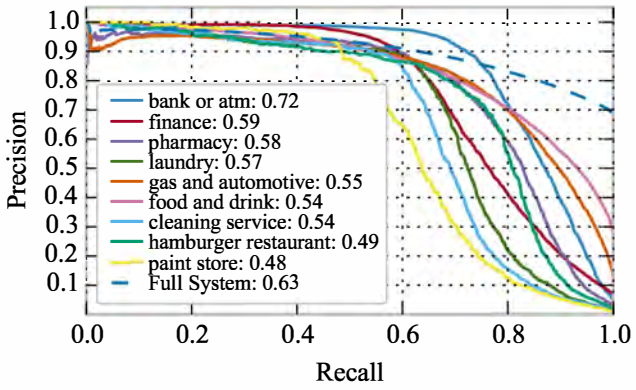



Figure 7. Precision recall curves for some of the top performing categories. The precision curve of the full system is shown as a dashed line. Recall at 90% precision is shown in the legend.

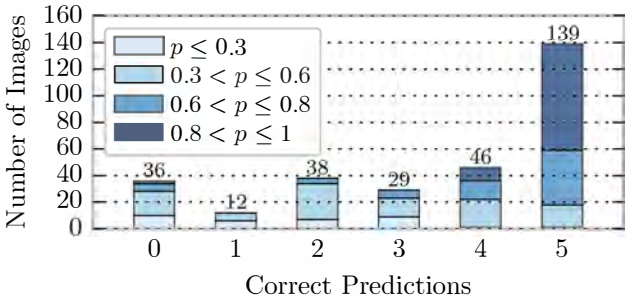



Figure 8. Histogram of correct labels in the top 5 predictions for a set of 300 manually verified images. Color indicates mean prediction confidence. Note that the confidence in prediction is strongly correlated with the accuracy.

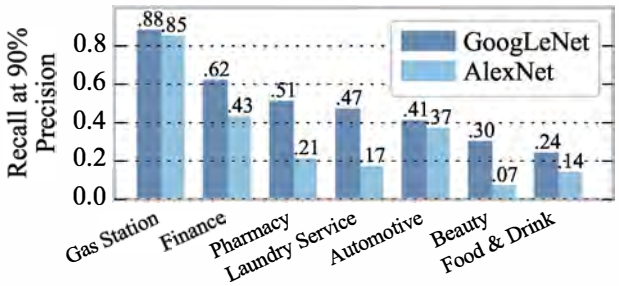



Figure 9. Recall at 90% precision for the GoogLeNet model used by our method and the AlexNet architecture [17] on a 13 category classification task. We show classes for which at least one of the models has recall $\geq 0.1$. GoogLeNet performs better on all categories. On some, such as PHARMACY recall more than doubles.

contain any of the true labels decreases with $k$. In order to save space we sum up all the probabilities for $k \in [15, 208]$ in one bin.

As mentioned above, one important metric for evaluation is recall at specific operating points. Figure 7 shows precision recall curves for some of the top performing categories and summary curve of the full system (dashed). The precision and recall of the full system is calculated by using the top-1 prediction. For many categories we are able to recover the majority of businesses while precision is held at 90%. For a classification system to be useful in a practical setting the classifier's returned confidence must be well correlated with the quality of its prediction. Figure 8 shows a histogram of the number of correctly predicted labels in the top 5 predictions on a set of images whose labels were manually verified. The mean prediction confidence is indicated by color intensity (darker means higher confidence). Note the strong correlation between confidence and accuracy; for confidence above 80% normally at least 4 of the top labels are correct.

The GoogLeNet network [27] incorporates a number of new design elements that make it particularly appealing: by not using fully connected upper layers, and forcing the network to go through dimensionality reduction stages, the network has far fewer parameters while being much deeper than previous methods. We evaluate the use of this architecture by comparing it to the AlexNet network proposed by Krizhevsky et al. [17] on a 13 category, *single label* classification task. We select a set of useful categories that a user might search for, and train both networks to predict one label per image. Figure 9 shows recall at 0.9 precision on all categories for which at least one method has recall $\geq 0.1$. GoogLeNet outperforms the baseline for all categories, and for some, e.g. PHARMACY, recall more than doubles.

Figure 10 shows 30 sample images from the test set with their top 5 predictions. The model is able to classify these images with their high level categories and with fine grained, specialized classes. For example, the top-right image was classified by the model as: BEAUTY, BEAUTY SALON, COSMETICS, HEALTH SALON, and NAIL SALON.

### 6.2. Human Performance Studies

Our system needs to perform at human level accuracy. Otherwise it will require a human verification post process. To estimate human accuracy on this task we have conducted two large scale human performance studies that check the agreement of operator-provided labels for the same business. In our experiments, subjects were shown images of businesses and selected one of 13 options (12 categories, and an OTHER category that stands for "none of the above".) Categories were chosen based on their financial impact for a business listing creator. The categories are: FOOD STORE, RESTAURANT/CAFFE, FINANCE, PHARMACY, REAL ESTATE, GAS STATION, AUTOMOTIVE (other than gas stations), FASHION, LODGING, LAUNDRY, PLACE OF WORSHIP, BEAUTY. Note that the studies used full resolution images as opposed to the $256 \times 256$ images given to the algorithm. The first study had 73,272 images, each was shown to two operators. The operators agreed on 69% of image labels. In our second



Authorized licensed use limited to: Penn State University. Downloaded on June 13,2024 at 18:51:02 UTC from IEEE Xplore. Restrictions apply.

| Operator Agreement | Number of images *Study 1* | *Study 2* |
|---|---|---|
| 100% | 50,425 | 9,938 |
| 75% | - | 9 |
| 66% | - | 8,535 |
| 50% | - | 133 |
| 0% | 22,847 | 1,300 |
| **Average Agreement** | **69%** | **78%** |

Table 1. Human Performance studies. In two large scale human studies we have found that manual labelers agree on a label for 69% and 78% of the images.

study we had 20,000 images, but each image was shown to *three* to *four* subjects. We found that the average agreement of the human operators was 78%. Table 1 shows a detailed summary of the human study results.

### 6.3. Analysis: Learning to Read With Weak Labels

For some of the images in Figure 10, such as the image of the dental center (top row, second image from right) it is surprising that the model was capable of classifying it correctly. It is hard to think of a "canonical" dental center look, but even if we could, it doesn't seem likely that this image would be it. In fact, without reading the text, it seems impossible to correctly classify it. This suggests that the system has learned to use text when needed. Figure 11 shows images from Figure 10 for which we have manually blurred the discriminative text. Note especially the image of the dental center, and of the auto dealer. After blurring the test "Dental" the system is confused about the dental center; it believes it is a beauty salon of some sorts. However, for the auto dealer, it is still confident that this is a place that sells things, and is related to transportation and automotive. To take this experiment to its logical extreme, we also show a synthetic image, which contains only the word *Pharmacy*. The classifier predicts the relevant labels for it.

To us this is a compelling demonstration of the power of Deep Convolutional Networks to learn the correct representation for a task. Similar to a human in a country in which she does not know the language, it has done the best it can – learn that some words are correlated with specific types of businesses. Note that it was never provided with annotated text or language models. It was only provided with what we would consider as *weak textual labels*, images that contain text and labeled with category labels. Furthermore, when text is not available the system is able to make an accurate prediction if there is distinctive visual information.

## 7. Discussion

Business category classification, is an important part of location aware search. In this paper we have proposed a method for fine grained, multi label, classification of business storefronts from street level imagery. We show that our system learned to extract and associate text patterns in multiple languages to specific business categories without access to explicit text transcriptions. Moreover, our system is robust to the absence of text, and when distinctive visual information is available, it is able to make correct predictions. We show our system achieves human level accuracy.

Using an ontology of entities with geographical attributes, we propagate label information during training, and produce a large set of 1.3 million images for a fine grained, multi label task. The use of non visual information, such as an ontology, to "ground" image data to real world entities is an exciting research direction, and there is much that can be done. For example, propagating information using the ontology at test time can increase both accuracy and recall. Node similarity in the ontology can be used to guide feature sharing between classes, and improve performance for seldom viewed classes.




Figure 11. A number of images from Figure 10 with the discriminative text in the image blurred (noted above the image). For some images, without the discriminative word the algorithm is confused (left column). For example, for the dental center, without the word *dental* it predicts a beauty salon. For other images, there is enough non textual information for the algorithm to be confident of the business category even when the text is blurred, for example the car dealership. Note the image of the nail spa: when the word *nail* is blurred the classifier falls back to more generic classes that fit the visual information - beauty salon, cosmetics, etc. As a final indicator to the ability of the network to learn textual cues we show a synthetic image where the only visual information is the word *pharmacy*. The network predicts relevant labels.



Authorized licensed use limited to: Penn State University. Downloaded on June 13,2024 at 18:51:02 UTC from IEEE Xplore. Restrictions apply.



| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| finance | .997 | shopping | .813 | prof. services | .998 | automotive | .999 | health & beauty | .992 | beauty | .997 |
| bank or atm | .994 | store | .805 | real estate agency | .995 | gas & automotive | .999 | health | .985 | beauty salon | .997 |
| atm | .976 | *construction* | *.662* | real estate | .992 | shopping | .999 | doctor | .961 | cosmetics | .995 |
| user op machine | .975 | home goods (s) | .530 | rental | .453 | store | .999 | emergency services | .960 | health salon | .994 |
| bank | .948 | *building material (s)* | *.300* | *finance* | .085 | vehicle dealer | .998 | dentist | .945 | nail salon | .953 |



| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| telecommunication | .826 | shopping | .923 | shopping | .920 | laundromat | .934 | food & drink | .947 | automotive | .999 |
| cell phone (s) | .796 | store | .908 | store | .916 | cleaners | .795 | food | .867 | gas & automotive | .999 |
| shopping | .627 | food & drink | .860 | sporting goods (s) | .625 | prof. services | .732 | restaurant or cafe | .722 | repairs | .999 |
| store | .627 | food | .849 | sports | .600 | laundry | .679 | restaurant | .621 | prof. services | .999 |
| *health & beauty* | *.116* | butcher shop | .824 | *textiles* | *.374* | cleaning service | .669 | beverages | .441 | car repair | .998 |



| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| shoe store | 1.00 | car repair | 1.00 | cafe | 1.00 | food & drink | 1.00 | liquor store | 1.00 | food & drink | .999 |
| shoes | 1.00 | gas & automotive | 1.00 | beverages | 1.00 | food | 1.00 | beverages | .999 | food | .998 |
| store | 1.00 | automotive | 1.00 | restaurant or cafe | 1.00 | restaurant or cafe | .999 | shopping | .998 | restaurant or cafe | .884 |
| shopping | 1.00 | prof. services | 1.00 | food & drink | 1.00 | restaurant | .999 | store | .998 | restaurant | .995 |
| clothing | .001 | repairs | 1.00 | food | 1.00 | hamburger restaurant | .936 | food & drink | .700 | fast food restaurant | .884 |



| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| health | .999 | prof. services | .999 | prof. services | .995 | *food & drink** | *.996* | food & drink | .825 | shopping | .932 |
| health & beauty | .999 | real estate | .996 | company | .982 | *food** | *.959* | food | .762 | store | .920 |
| pharmacy | .997 | real estate agency | .973 | cleaning service | .975 | *restaurant** | *.931* | restaurant or cafe | .741 | florist | .896 |
| emergency services | .996 | *rental* | *.132* | laundry | .970 | *restaurant or cafe** | *.909* | restaurant | .672 | *fashion* | *.077* |
| shopping | .989 | *consultant* | .029 | dry cleaner | .966 | *asian** | *.647* | beverages | .361 | *gift shop* | *.071* |



| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| prof. services | .594 | gas station | .996 | shopping | .489 | shopping | .719 | beauty | .999 | place of worship | .990 |
| legal services | .346 | transportation | .996 | store | .467 | store | .713 | health & beauty | .999 | church | .988 |
| lawyer | .219 | gas & automotive | .995 | *prof. services* | *.289* | home goods (s) | .344 | cosmetics | .998 | *education/culture** | *.031* |
| *insurance* | *.129* | *government* | *.001* | *services* | *.246* | furniture store | .299 | health salon | .998 | *assoc./organization** | *.029* |
| *insurance agency* | *.103* | *gastronomy* | *.001* | *gas & automotive* | *.219* | mattress store | .240 | nail salon | .949 | *prof. services* | *.027* |

Figure 10. Top predictions for sample images from the test set. Predictions marked in *red* disagree with ground truth labels; in some cases the classifier is correct and the ground truth label is wrong (marked with *). For example, see asian restaurant (fourth image from left, row before last), for which the top 5 predictions are all correct, but do not agree with the ground truth. The mark (s) is an abbreviation for store.



Authorized licensed use limited to: Penn State University. Downloaded on June 13,2024 at 18:51:02 UTC from IEEE Xplore. Restrictions apply.

## Acknowledgments

The authors wish to thank Ian Goodfellow, Christian Szegedy, and Yaser Sheikh for helpful discussions.

## References

[1] D. Anguelov, C. Dulong, D. Filip, C. Frueh, S. Lafon, R. Lyon, A. Ogale, L. Vincent, and J. Weaver. Google street view: Capturing the world at street level. *Computer*, 2010. 2

[2] S. Arora, A. Bhaskara, R. Ge, and T. Ma. Provable bounds for learning some deep representations. *arXiv:1310.6343 [cs, stat]*, oct 2013. 4

[3] J. Bhogal, A. Macfarlane, and P. Smith. A review of ontology based query expansion. *Information processing & management*, 43(4):866–886, 2007. 4

[4] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *European Conference of Computer Vision (ECCV)*. 2010. 2

[5] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *British MAchine Vision Conference (BMVC)*, 2014. 5

[6] N. Cornelis, B. Leibe, K. Cornelis, and L. Van Gool. 3d urban scene modeling integrating recognition and reconstruction. *International Journal of Computer Vision*, 2008. 2

[7] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. aurelio Ranzato, A. Senior, P. Tucker, K. Yang, Q. V. Le, and A. Y. Ng. Large scale distributed deep networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1223–1231. 2012. 2, 4

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html. 2

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2010. 2

[10] A. Flores and S. Belongie. Removing pedestrians from google street view images. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010. 2

[11] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4), 1980. 2

[12] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks. *arXiv preprint arXiv:1312.6082*, 2013. 2, 5

[13] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. *arXiv preprint arXiv:1302.4389*, 2013. 2

[14] A. B. Hillel and D. Weinshall. Subordinate class recognition using relational object models. *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, 2007. 2

[15] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012. 2

[16] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009. 2

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 2012. 2, 6

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 1998. 2

[19] Y. J. Lee, A. A. Efros, and M. Hebert. Style-aware mid-level representation for discovering visual connections in space and time. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, 2013. 2

[20] M. Lin, Q. Chen, and S. Yan. Network in network. *CoRR*, abs/1312.4400, 2013. 4

[21] B. Micusik and J. Kosecka. Piecewise planar city 3d modeling from street view panoramic sequences. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009. 2

[22] P. X. Nguyen, K. Wang, and S. Belongie. Video text detection and recognition: Dataset and benchmark. In *Winter Conference on Applications of Computer Vision (WACV)*, Steamboat Springs, CO, March 2014. 3

[23] N. F. Noy. Semantic integration: a survey of ontology-based approaches. *ACM Sigmod Record*, 33(4):65–70, 2004. 4

[24] H. Parsa, J. T. Self, D. Njite, and T. King. Why restaurants fail. *Cornell Hotel and Restaurant Administration Quarterly*, 46(3):304–322, 2005. 1

[25] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. *CoRR*, 2014. 5

[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge, 2014. 2

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv:1409.4842 [cs]*, sep 2014. 2, 4, 6

[28] L. Vincent. Taking online maps down to street level. *Computer*, 2007. 2

[29] C. Wah, S. Branson, P. Perona, and S. Belongie. Multiclass recognition and part localization with humans in the loop. In *IEEE International Conference on Computer Vision (ICCV)*, 2011. 2

[30] K. Wang and S. Belongie. Word spotting in the wild. In *European Conference on Computer Vision (ECCV)*, Heraklion, Crete, Sept. 2010. 3



Authorized licensed use limited to: Penn State University. Downloaded on June 13,2024 at 18:51:02 UTC from IEEE Xplore. Restrictions apply.

[31] J. Xiao and L. Quan. Multiple view semantic segmentation for street view images. In *Computer Vision, 2009 IEEE 12th International Conference on*, 2009. 2

[32] B. Yao, G. Bradski, and L. Fei-Fei. A codebook-free and annotation-free approach for fine-grained image categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2

[33] B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 2

[34] A. R. Zamir and M. Shah. Accurate image localization based on google maps street view. In *Computer VisionECCV 2010*. Springer, 2010. 2

[35] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional neural networks. *arXiv preprint arXiv:1311.2901*, 2013. 2



Authorized licensed use limited to: Penn State University. Downloaded on June 13,2024 at 18:51:02 UTC from IEEE Xplore. Restrictions apply.

**Appendix VAROL01**

```
001    993213256978014 51
leader 03570cam a2200577 i 4500
005    20240212213017.0
006    m     o  d
007    cr mn|||||||||
008    150408t20152015swaua    ob    100 0 eng d
050    #4$aTA1637 $b.I543 2014e
040    ##$aCOO $beng $erda $cCOO $dGAT $dNHM $dOCLCO $dCUS $dOCLCO $dOCLCF $dOCLCO $dCRU $dLVT $dCAUOI $dOCLCO $dOCLCQ $dOCLCO $dOCLCA $dOCLCL $dOCLCQ
020    ##$z9781628415582
020    ##$z1628415584
049    ##$aMAIN
111    2#$aInternational Conference on Graphic and Image Processing $n(6th : $d2014 : $cBeijing, China)
264    #1$aBellingham, Washington : $bSPIE, $c[2015]
264    #4$c©2015
245    10$aSixth International Conference on Graphic and Image Processing (ICGIP 2014) : $b24-26 October 2014, Beijing, China / $cYulin Wang, Xudong Jiang, David Zhang, editors ; organized by IACSIT--International Association of
336    ##$atext $btxt $2rdacontent
337    ##$acomputer $bc $2rdamedia
338    ##$aonline resource $bcr $2rdacarrier
300    ##$a1 online resource : $billustrations (some color).
347    ##$atext file $bPDF $2rda
490    1#$aProceedings of SPIE, $x0277-786X ; $vvolume 9443
588    0#$aOnline resource; title from PDF title page (SPIE digital library, viewed September 18, 2015).
504    ##$aIncludes bibliographical references.
650    #0$aImage processing $xDigital techniques $vCongresses.
650    #0$aImaging systems $xImage quality $vCongresses.
650    #0$aImage processing $vCongresses.
650    #6$aTraitement d'images $xTechniques numériques $vCongrès.
650    #6$aImagerie (Technique) $xQualité de l'image $vCongrès.
650    #6$aTraitement d'images $vCongrès.
650    #7$aImage processing $2fast
650    #7$aImage processing $xDigital techniques $2fast
650    #7$aImaging systems $xImage quality $2fast
655    #7$aConference papers and proceedings $2fast
710    2#$aSPIE (Society), $eissuing body, $esponsoring body.
776    08$iPrint version: $z9781628415582 $w(OCoLC)921269938
758    ##$ihas work: $aSixth International Conference on Graphic and Image Processing (ICGIP 2014) (Text) $1https://id.oclc.org/worldcat/entity/E39PCFQ7HKtccFmXv4Qtdvxjyd $4https://id.oclc.org/worldcat/ontology/hasWork
700    1#$aWang, Yulin $c(College teacher), $eeditor. $1https://id.oclc.org/worldcat/entity/E39PCjHCTHgbtH3TR8GFR7dJDq
700    1#$aJiang, Xudong. $uNanyang Technological University, Beijing (China), $eeditor.
700    1#$aZhang, David, $d1949- $eeditor. $1https://id.oclc.org/worldcat/entity/E39PBJfx6DmMJxbGyY3mmcmDbd
830    #0$aProceedings of SPIE--the International Society for Optical Engineering ; $vv. 9443. $x0277-786X
856    40$3SPIE Digital Library $uhttp://proceedings.spiedigitallibrary.org/volume.aspx?volume=9443&issue=1
856    40$zSPIE. Restricted to UC campuses $uhttp://uclibs.org/PID/281717
856    40$uhttp://liverpool.idm.oclc.org/login?url=http://proceedings.spiedigitallibrary.org/volume.aspx?volume=9443 $z<img src="/screens/gifs/go4.gif" alt="Go button" border="0" width="21" height="21" hspace="7" align="middle">
856    40$uhttps://login.uproxy.library.dc-uoit.ca/login?qurl=https://www.spiedigitallibrary.org/conference-proceedings-of-spie/9443 $zeBook available for UOIT via SPIE. Click link to access
994    ##$a92 $bOCACL
035    ##$a(OCoLC)907375748
035    ##$a(OCoLC)ocn907375748
```

CONFERENCE PROCEEDING

# Sixth International Conference on Graphic and Image Processing (ICGIP 2014) : 24-26 October 2014, Beijing, China

International Conference on Graphic and Image Processing (6th : 2014 : Beijing, China)
2015; Bellingham, Washington : SPIE

Online access

TOP
SEND TO
VIEW ONLINE
ITEM DETAILS
LINKS

## Send to

E-MAIL PERMALINK CITATION PRINT ENDNOTE EASYBIB
EXPORT RIS BIBTEX EXPORT TO EXCEL

## View Online

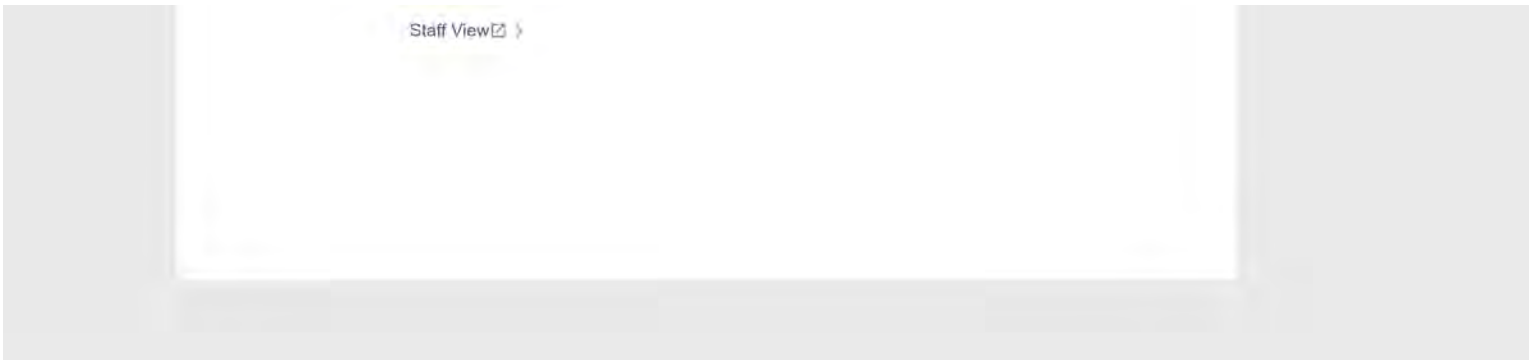
**Full text availability**

SPIE Digital Library (Proceedings) [SHOW LICENSE]
*Access for all on campus; login required from off-campus.*
*Details about who can access online resources*

## Item Details

| | |
|---|---|
| **Title** | Sixth International Conference on Graphic and Image Processing (ICGIP 2014) : 24-26 October 2014, Beijing, China |
| **Author** | International Conference on Graphic and Image Processing (6th : 2014 : Beijing, China)<br>Wang, Yulin (College teacher), editor.<br>Jiang, Xudong. Nanyang Technological University, Beijing (China), editor.<br>Zhang, David, 1949- editor.<br>SPIE (Society), issuing body, sponsoring body. |
| **Attribution** | Yulin Wang, Xudong Jiang, David Zhang, editors ; organized by IACSIT--International Association of Computer Science and Information Technology ; sponsored by Wuhan University (China), IACSIT--International Association of Computer Science and Information Technology ; published by SPIE |
| **Subjects** | Image processing -- Digital techniques -- Congresses<br>Imaging systems -- Image quality -- Congresses<br>Image processing -- Congresses<br>Image processing<br>Image processing -- Digital techniques<br>Imaging systems -- Image quality |
| **Related Titles** | Proceedings of SPIE--the International Society for Optical Engineering ; v. 9443. 0277-786X |
| **Genre** | Conference papers and proceedings |
| **Language** | English |
| **Format** | 1 online resource : illustrations (some color). |
| **Technical Details** | Encoding format: PDF |
| **Description** | **Series statement:** Proceedings of SPIE, 0277-786X ; volume 9443<br>Includes bibliographical references. |
| **Publisher** | Bellingham, Washington : SPIE |
| **Publication Date** | 2015 |
| **Identifier** | ISBN: 9781628415582 (Additional Physical Format) |
| **OCLC Number** | 907375748 |
| **Alma IZ Record Number** | 99161902534101452 |
| **Alma NZ Record Number** | 99321325697801451 |

## Links


Staff View

**Appendix YEH01**


Page 149 of 169

# MARC View

| | | | | |
|---|---|---|---|---|
| LEADER 01328nam a22002652 4500 | | | | |
| 001 20030141 | | | | |
| 003 SIRSI | | | | |
| 005 20170314130250.0 | | | | |
| 006 m d | | | | |
| 007 cr n | | | | |
| 008 040903e200401uucau s\|\|\|\|\|\|\| 2\|eng\|d | | | | |
| 020 | | | a\| 9780769521589 | |
| 020 | | | a\| 0769521584 (Trade Paper, CD-ROM, Mixed Media) c\| USD 445.00 Retail Price (IEEE) 9\| Active Record | |
| 037 | | | a\| PR2158 b\| 00029433 | |
| 040 | | | a\| BIP US d\| WaSeSS | |
| 110 | 2 | | a\| Institute of Electrical and Electronics Engineers, Inc. Staff e\| Author | |
| 245 | 1 | 0 | a\| Computer Vision and Pattern Recognition, CVPR 2004 h\| [electronic resource]: b\| Proceedings, 22nd Conference, Washington, D.C. | |
| 260 | | | a\| Los Alamitos : b\| IEEE Computer Society Press c\| Jan. 2004 a\| Piscataway : b\| IEEE [Distributor] | |
| 300 | | | a\| 2010 p. | |
| 506 | | | a\| License restrictions may limit access. | |
| 521 | | | a\| Scholarly & Professional b\| IEEE Computer Society Press | |
| 655 | | 0 | a\| Electronic books. | |
| 856 | 4 | 0 | u\| http://SK8ES4MC2L.search.serialssolutions.com/?sid=sersol&SS_jc=TC0000506958&title=Computer%20Vision%20and%20Pattern%20Recognition%2C%20CVPR%202004%3A%20Proceedings%2C%2022nd%20Conference%2C%20Washington%2C%20D.C | |
| 955 | | | a\| Books In Print, (c) 2011 R.R. Bowker LLC | |
| 949 | | | a\| Electronic resource w\| ASIS c\| 1 i\| 20030141-1001 l\| ONLINE m\| ONLINE r\| Y s\| Y t\| ONLINE u\| 3/21/2017 o\| .STAFF. 360 MARC | |

ASK A LIBRARIAN

PennState

Copyright @2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

Legal Statements | PSU Hotlines

PENN STATE UNIVERSITY LIBRARIES

Libraries Home (814) 865-6368

Libraries Intranet (Staff Only)

Accessibility Help

Website Feedback

Policies and Guidelines

Staff Directory

CONNECT WITH PENN STATE UNIVERSITY LIBRARIES

Facebook

Twitter

Instagram

# Computer Vision and Pattern Recognition, CVPR 2004 [electronic resource]: Proceedings, 22nd Conference, Washington, D.C.

Corporate Author: Institute of Electrical and Electronics Engineers, Inc. Staff

Published: Los Alamitos : IEEE Computer Society Press Jan. 2004 Piscataway : IEEE [Distributor]

Physical Description: 2010 p.

ASK A LIBRARIAN

Access Online

serialssolutions.com

**Restrictions on Access:**
License restrictions may limit access.

**Genre(s):**
Electronic books

**ISBN:**
9780769521589
0769521584 (Trade Paper, CD-ROM, Mixed Media)

**Audience Notes:**
Scholarly & Professional IEEE Computer Society Press

View MARC record | catkey: 20030141

PennState

Copyright @2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

Legal Statements | PSU Hotlines

PENN STATE UNIVERSITY LIBRARIES

Libraries Home (814) 865-6368
Libraries Intranet (Staff Only)
Accessibility Help
Website Feedback
Policies and Guidelines
Staff Directory

CONNECT WITH PENN STATE UNIVERSITY LIBRARIES

Facebook
Twitter
Instagram

IEEE.org IEEE *Xplore* IEEE SA IEEE Spectrum More Sites Donate Cart Create Account Personal Sign In

Browse My Settings Help

Access provided by: Penn State University Sign Out

Access provided by: Penn State University Sign Out

All

ADVANCED SEARCH



# Searching the Web with mobile images for location recognition

**Publisher: IEEE**

Cite This PDF

T. Yeh ; K. Tollmar ; T. Darrell All Authors

62 Cites in Papers

122 Cites in Patents

540 Full Text Views

Alerts

Manage Content Alerts

Add to Citation Alerts

Abstract

Document Sections

1. Introduction
2. Previous Work
3. Image-Based Web Search
4. Experiments
5. Conclusions and Future Works

Authors

Figures

References

Citations

Keywords

Metrics

Loading math: 81%

Downl PDF

**Abstract:**
We describe an approach to recognizing location from mobile devices using image-based Web search. We demonstrate the usefulness of common image search metrics applied on ... **View more**

**Metadata**

**Abstract:**
We describe an approach to recognizing location from mobile devices using image-based Web search. We demonstrate the usefulness of common image search metrics applied on images captured with a camera-equipped mobile device to find matching images on the World Wide Web or other general-purpose databases. Searching the entire Web can be computationally overwhelming, so we devise a hybrid image-and-keyword searching technique. First, image-search is performed over images and links to their source Web pages in a database that indexes only a small fraction of the Web. Then, relevant keywords on these Web pages are automatically identified and submitted to an existing text-based search engine (e.g. Google) that indexes a much larger portion of the Web. Finally, the resulting image set is filtered to retain images close to the original query. It is thus possible to efficiently search hundreds of millions of images that are not only textually related but also visually relevant. We demonstrate our approach on an application allowing users to browse Web pages matching the image of a nearby location.

**Published in:** Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.

**Date of Conference:** 27 June 2004 - 02 July 2004 **DOI:** 10.1109/CVPR.2004.1315147

More Like This

**Date Added to IEEE *Xplore*:** 19 July 2004

**Publisher:** IEEE

**Print ISBN:**0-7695-2158-4

**Conference Location:** Washington, DC, USA

**Print ISSN:** 1063-6919

Contents

## SECTION 1.
# Introduction

Content-based image retrieval (CBIR) has been an ongoing research topic for the past decade. One of the difficulties reported by users is that queries are hard to specify with the conventional input modalities; unlike keywords, a visual query may require the user to sketch or select parameter values for each low-level feature [10]. For recognizing realworld entities, searching with an actual image seems ideal. The advent of a new generation of camera-equipped phones and PDAs provides a unique platform for image-based web search.

We apply this paradigm to location recognition. Location recognition with radio or other signals (e.g., GPS, RFID tags, etc.) is possible, but an image-based approach can recognize locations distinct from a device's actual position. (e.g., “Tell me about that building over there.”, while taking a snapshot of the building.) We have found that many landmarks and other prominent locations already have images on relevant web pages, so matching the mobile image to web documents can provide a reasonable location cue. We take images of an unknown location from a mobile camera, send them to a server for processing, and return a set of matching web pages.

Content-based search over the entire web is a computationally prohibitive proposition. Searching an image database on a scale similar to any commercial keyword-based image search engine, such as Google with 425 million images, remains a research challenge. We propose a hybrid approach that can provide content-based image retrieval by combining both image-based and text-based search. We leverage the fact that there will be redundant landmark images on the web and that a search in a subset of the web will likely find a matching image.

Our approach is to first search a bootstrap set of images obtained by web-crawling a restricted domain. The domain is selected based on the expected application, for example tourism-related sites for a particular geographic location. To recover relevant pages across the full web, we exploit a keyword-based search followed by a content-based filtering step. Keywords are extracted from web pages with matching images in the bootstrap set. Instead of running CBIR over hundreds of millions of images, we only need to operate on a seed set of images and on the images returned from the keyword-based search (see Figure 1).

In this paper, we focus on one particular class of images-landmarks for recognizing location. The goal is to recognize the location by finding web pages with an image matching the image of the location. In the next section we will review relevant related work. Following that, we present our approach to keyword and image-based web search and detail the component modules of our method. We present experiments on a tourist location recognition task to evaluate the components of our approach. We conclude with summary of our contribution.

## SECTION 2.
# Previous Work

Typesetting math: 81%

Most commercially successful image search engines are text-based. Corbis features a private database of millions of high-quality photographs and artwork that are manually tagged with keywords and organized into categories. Google has indexed more than 425 million web pages and inferred their content in the form of keywords by analyzing the text on the page adjacent to the image, the image caption, and other text features. In both cases, the image search engine searches for images based on text keywords. Since the visual content of the image is ignored, images that are visually unrelated can be returned in the search result. However, this approach has the advantage of text search-semantically intuitive, fast, and comprehensive.

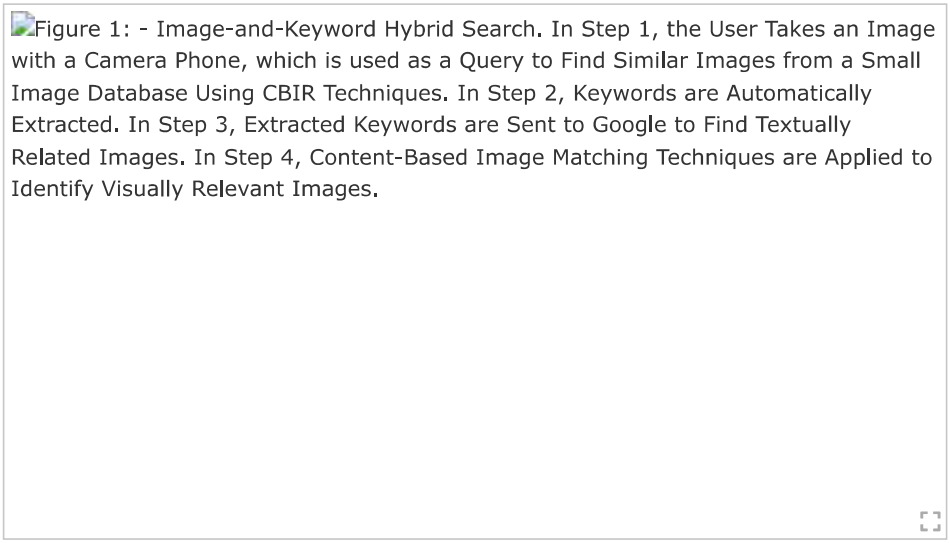



**Figure 1:** Image-and-Keyword Hybrid Search. In Step 1, the User Takes an Image with a Camera Phone, which is used as a Query to Find Similar Images from a Small Image Database Using CBIR Techniques. In Step 2, Keywords are Automatically Extracted. In Step 3, Extracted Keywords are Sent to Google to Find Textually Related Images. In Step 4, Content-Based Image Matching Techniques are Applied to Identify Visually Relevant Images.

Searching for images from images is often called content - based image retrieval and has been an ongoing research topic for many years [10]. Although the algorithms for analyzing image content in general run significantly slower than those for analyzing text, the search result can exhibit stronger visual relevance; we can have greater confidence as to whether the images retrieved fit the intended query. One of the first such systems was IBM's Query-By-Image-Content (QBIC) system [3]. It supported search by example images, user-drawn pictures, or selected color and texture patterns, and was applied mainly to custom, special-purpose image databases. In contrast, the Webseek system [11] searched generically on the World Wide Web for images. This system incorporated both keyword-based and content-based techniques; the keyword search returned a set of images among which users could further search by color histogram similarity with relevance feedback. The Diogenes system used a similar bi-modal approach for searching images of faces on the web [2]. A face detector operated as a screening process to filter out nonface images from the set of initial images returned by a keyword-based query. Then, a face recognition module processed the remaining images to identify particular faces.

Describing what images to search for can be tricky. QBIC lets users choose the relative weights of each feature in determining image similarity. Lew et. al proposed a system that allows users to sketch the outline of the image [6]. VisualSeek features an interface to paint a color template for template-based image matching [11]. However, these systems failed to provide an intuitive interface. The correspondence between low-level parameter values and high-level visual concept is not so obvious to ordinary users. Drawing can require too much effort and artistic skill. One solution is the search-by-example approach. For example, ImageRover presents the user with a set of starting images as examples and uses a “relevance feedback” framework to refine the search iteratively based on what the user says about the relevance of each image [9].

Due to the complexity typically involved in CBIR algorithms, there is a tradeoff between interactivity and comprehensiveness. Many existing CBIR systems use simple nearest-neighbor techniques to find matching images, which cannot scale to a large and comprehensive image

Typesetting math: 81%

database and still perform fast enough to be interactive. Hierarchical indexing with Self-Organizing Maps [5] and related techniques is possible, but the speed improvement is very limited. So far, practical CBIR systems operate on a scale much smaller than their text-based counterparts. We were not aware of any CBIR system that had gone beyond the million images mark in scale [6], [11], [9], [2], [3].

Our approach to image-based location-recognition is closely related to efforts in the wearable-computing and robotics literature on navigation and scene recognition. The wearable museum guiding system built by Starner and colleagues uses a head-mounted camera to record and analyze the visitor's visual environment [12]. Computer vision techniques based on oriented edge histograms were applied to recognize objects in the field of view. Based on the objects seen, the system estimates the location in the museum and displays relevant information. The focus of this system was on recalling prior knowledge of locations-which item is exhibited where-rather than finding new information about locations. Torralba et al. generalized this type of technique and combined it with a probabilistic model of temporal dynamics, so that information from a video sequence from a wearable device could be used for mapping and navigation tasks [13]. In these robotics and wearable computing systems, recognition was only possible in places where the system had physically been before. In our system location-relevant information can be found from a single image of a novel place.

## SECTION 3.
# Image-Based Web Search

Web authors tend to include semantically related text and images on web pages. There exists a high correlation between semantic relevancy and spatial proximity that can be exploited on the web; pieces of knowledge close together in cyberspace tend to be also mutually relevant in meaning [1]. To find information about a well-known landmark, we can search for web pages with images matching the image of the current location, and analyze the surrounding text. Using this approach, the location-recognition problem can be cast as a CBIR problem-if methods can be found to match mobile images to the web despite time, pose, and weather variation, they can serve as a useful tool for mobile web search (and in the particular application we consider here, location-based computing.)

With a camera phone, a user can take an picture of a prominent landmark around the current location. This picture can be used as a query to find images similar to this landmark on the web using content-based image matching techniques. Relevant keywords can be found in the surrounding text and used directly as a location context cue, or used for further interactive browsing to find relevant information resources. A very attractive property of this method is that it requires no special-purpose communications infrastructure or prior location database, and that from a user-interface perspective users can specify nearby locations with a very natural interface action-taking a picture of the intended place.

For a pure CBIR system to match against the millions of images on the web in real-time is currently computationally infeasible. However, using a hybrid keyword-and-image query system, we effectively implement CBIR over the entire 425 million images without having to apply a content-based metric on every single image. Such a hybrid design benefits from the power of both keyword-based search (speed and comprehensiveness) and image-based search (visual relevancy).

We leverage an existing keyword-based image search engine, Google, which has indexed more than 425 million images. We extract keywords from web pages found in a content-based search in a bootstrap database, and use these keywords on Google to search its larger database of images for images we want. Recall that one shortcoming of keyword-based search is the existence of visually unrelated images in the result set. To overcome this, we apply a filtering step, where we run CBIR on this small set of images to identify visually related images. In this way we retrieve images that are not only visually relevant but also textually related.

### 3.1. Image Matching Metrics

Typesetting math: 81%

Having the right feature set and image representation is very crucial for building a successful CBIR system. The performance of general object matching in CBIR systems is far from perfect-image segmentation and viewpoint variation are significant problems. Fortunately, finding images of landmarks requires analysis over the entire image, making general image segmentation unnecessary. Also, users ask about a location most likely because they are physically there and there are a much smaller number of physically common viewpoints of prominent landmarks than in the entire viewsphere of a common object. Consequently, the simplicity of recognizing location images naturally makes it an attractive test-bed to test the idea of an image-and-text hybrid search approach.

We have experimented with two image-matching metrics on the task of matching mobile location images to images on the World Wide Web. The first metric is based on the *energy spectrum*, the squared magnitude of the windowed Fourier transform of an image. Given an image i, the discrete Fourier transform can computed asI({f_x},{f_y}) = \sum\limits_{x,y = 0}^{N - 1} i (x,y)h(x,y) {e^{- j2\pi ({f_x}x + {f_y}y))}}

View Source

where h(x, y) is a circular hanning window to reduce the boundary effect. The amplitude component of the above term represents the spatial frequency spread everywhere in the image. It embeds unlocalized information about the image structure such as orientation, smoothness, length, and width of the contours that compose the entire scene in the image. This type of representation was demonstrated by [13] to be invariant to object arrangement and object identity.

The second metric is based on wavelet decompositions. We compute the local texture features of an image I by convolving it with steerable filters [4] over 2 different scales on the intensity (grayscale) image to obtain a representation as follows:\lambda=G_{\theta_{i}}(S_{j}(I))

View Source

where G is the steerable filter for 6 different \theta {\rm s}, evenly spaced between 0 and 2\pi, and S is the scaling operator. This can produce 6 \times 2=12 subbands. Since this gives us only the local representation of the image, we take the mean values of the magnitude of the local features averaged over large windows to capture global image propertiesm(x) = \sum\limits_{{x^\prime}} {|\lambda (i)|} \cdot w({x^\prime} - x)

View Source

where w is the averaging window. The final representation is downsampled to have a spatial resolution of 4 by 4 pixels resulting in a dimensionality of 4\times 4\times 12=96.

Given a mobile image of some landmark, similar images can be retrieved by finding the k nearest neighbors in the database using either of the two metrics (e.g. see Figure 2, where k=16). However, the high dimensionality (d) of the feature involved in the metric can be problematic. To reduce the dimensionality, we computed principal components over a large number of landmark images on the web. Then, each feature vector can be projected onto the first 32 (a number picked empirically) principal components which form the final feature vector for the images.

### 3.2. Extracting Useful Textual Information

After similar landmark images are found, we need to analyze the text of the web documents containing these images with the intention to uncover any useful information about this particular landmark. We devise a simple frequency-based approach to identifying named-entities (e.g. name, location, telephone number) by computing the *term frequency inverse document frequency* metric \mu for each bi-gram and trigram found in the document:\mu (w) = {{df(w)} \over {tf(w)}}

View Source

Typesetting math: 81%

where tf(w) is the global frequency of w over a large corpus and df(w) is the local frequency of w within the document [8]. The underlying intuition is that people tend to find the sudden occurrence (high df) of uncommon things (low tf) interesting. A good estimate of the global word frequency can be the number of links Google indexes for this word. The top n word combinations with the highest f can be reported as the potential named-entities to the users.

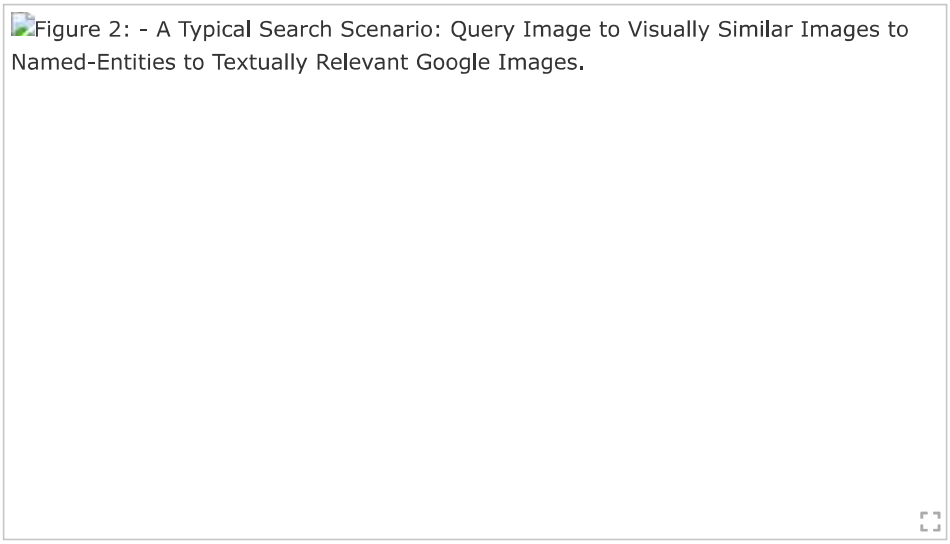



**Figure 2:** A Typical Search Scenario: Query Image to Visually Similar Images to Named-Entities to Textually Relevant Google Images.

Figure 2 illustrates a typical search scenario. A user initiates a search request by taking an image of some prominent landmark (e.g. MIT Dome) at the current location. Sixteen visually similar images to the query image are found in the web image database. Five of them are correct (1,3,4,9, and 14). The content of the source HTML documents of these 5 images are analyzed and potential named-entities are found and ranked. These named-entities can be submitted as keywords to a text-based search engine (e.g. Google) to find more images. However, it can not be guaranteed that these images are visually similar to the questioned landmark.

### 3.3. Content-Filtered Keyword Search

The name-entities extracted from the web pages can provide the users with not only the basic notion about this location but also the possibility to acquire further information from a powerful text-based search engine (e.g. Google).

One possibility is to search for more web pages that might lack any visual content but nevertheless indirectly share similarity with the query image through text. We can describe such similarity inference notationally:{I_q}\mathop \leftrightarrow \limits_{image} {I_v},{P_v}\mathop \leftrightarrow \limits_{text} \,{P_t} \Rightarrow {I_q} \leftrightarrow {P_t}

View Source

where I_{q} denotes the query image, I_{v} denotes the found similar image, P_{v} denotes the web page containing I_{v}, P_{v} denotes a web page containing similar keywords as P_{t}, and \leftrightarrow denotes a similarity relation based on either text or image. Here, P_{t} would not have been found otherwise if either of the pure image-based search or pure text-based search were employed.

Another possibility is to use the keywords to search for images I_{t}. The similarity inference can be denoted as:{I_q}\mathop \leftrightarrow \limits_{image} {I_v},{P_v}\mathop \leftrightarrow \limits_{text} {I_t} \Rightarrow {I_q} \leftrightarrow {I_t}.

View Source

Typesetting math: 81%

While I_{q} and I_{t} are textually related, to guarantee visual similarity, a content-based filtering step needs to be applied to keep keep only those images visually close to the query image. As shown in Figure 4, a search on the query image I_{q} (a 15 floor building) yields the keywords “MIT Green Building”. These keywords are submitted to Google to retrieve 16 such I_{t}{\rm s} that are textually relevant but not quite visually similar. Sorting each I_{t} according to its distance to I_{q}, we are able to boost the ranks of the visually similar images to the top.

Moreover, there might exist images visually distant but conceptually close to the query image; they can help us learn more about this location. We can use a bottom-up, opportunistic clustering technique that iteratively merges data points to uncover visually coherent groups of images. Initially, each image is considered as the center in its own cluster. At each iteration, two closest clusters, in terms of the same distance metric used in other parts of the system, are merged. The process will continue until the distance between every cluster pair has exceeded a certain threshold. If a cluster is reasonably large, it means the images in this group represent some potentially significant common concept. In the example shown in Figure 4, four images are removed from the result set because they can not be clustered with any other image.

## SECTION 4.
# Experiments

A series of experiments was conducted to evaluate the effectiveness of our algorithms for location recognition by matching mobile images to a web database. We chose the task of recognizing landmark images on a university campus. We ran a web crawler on *mit.edu* to collect images and their source URL's to build an image database as the bootstrap set. Images too small to contain interesting location information, less than 100 by 100, were discarded. The resulting database contained roughly 33,000 images. Since dealing with images of various dimensions and sizes can be tricky, we resized each image to 100 by 100 for uniformity.

To find similar landmark images, it would not be useful to search images that do not contain any landmark (e.g. faces, animals, or logos). Thus, an image classifier was used to classify these 33,000 images as landmark or nonland-mark. The nonlandmark images were then removed from the database to reduce the search space to 2000+ images. The image classifier was trained using a method similar to [7] that was demonstrated to be successful in classifying indoor-outdoor images by examining color and texture characteristics.

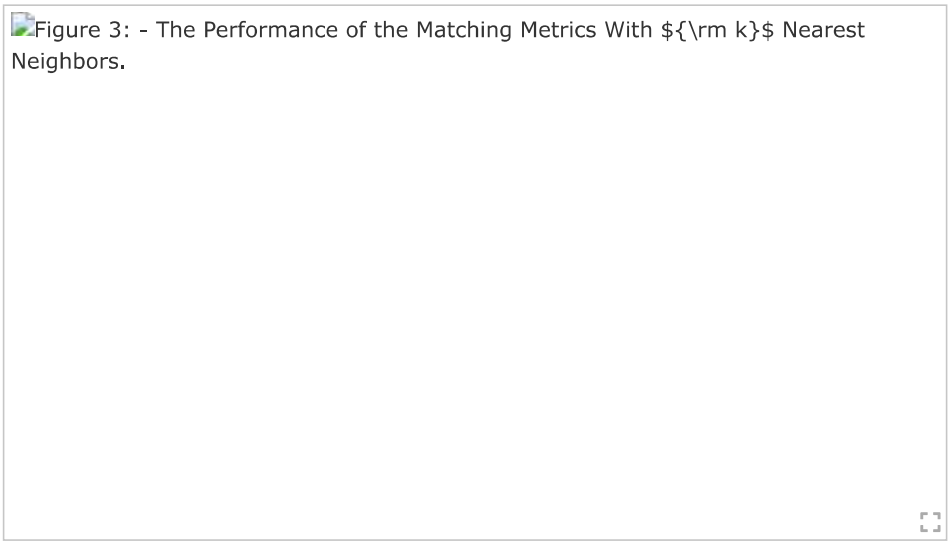



**Figure 3:** The Performance of the Matching Metrics With {\rm k} Nearest Neighbors.

Since we were concerned with location recognition in a mobile setting, the test queries consisted of landmark images acquired with a camera-equipped mobile device. We selected 5 prominent

Typesetting math: 81%

landmarks and asked volunteers to take 20 images of each of these landmarks using a Nokia 3650 camera phone. A total of 100 query images were obtained for 100 test runs.

The matching metrics discussed in section 3.1 were used in our experiments. For each query image, the k nearest images were retrieved from the database to form the search result. The same volunteers were asked to evaluate the search result. A success was defined as the occurrence of at least one correct landmark image among the first k nearest images. The findings are summarized in Figure 3, which plots the percentage of test cases that were successful over different k.

At k=16, the likelihood of finding correct images in the result set is reasonably high. This indicates that even with these standard feature sets it is possible to find similar location images for recognition tasks.

## SECTION 5.
# Conclusions and Future Works

In this paper, we explored the potential of web search with mobile images using both images and text. We have shown it is possible to conduct fast and comprehensive CBIR over hundreds of millions of images using a text-based search engine on keywords generated from an initial search. We presented an approach to recognizing location from mobile devices using image-based web search, and showed that common image search metrics can match images captured with a camera-equipped mobile device to images found on the

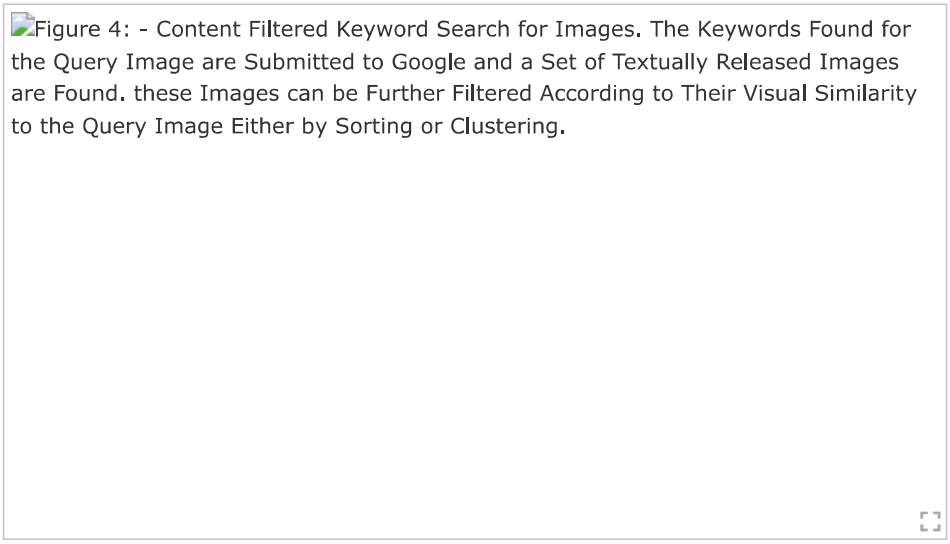



**Figure 4:** Content Filtered Keyword Search for Images. The Keywords Found for the Query Image are Submitted to Google and a Set of Textually Released Images are Found. these Images can be Further Filtered According to Their Visual Similarity to the Query Image Either by Sorting or Clustering.

World Wide Web or other general-purpose databases.

A hybrid image-and-keyword searching technique was developed that first performed an image-based search over images and links to their source web pages in a bootstrap database that indexes only a small fraction of the web and a simple frequency-based procedure was employed to extract relevant keywords from these web pages. These keywords can be submitted to an existing text-based search engine (e.g. Google) that indexes a much larger portion of the web. The resulting image set is then filtered to retain images close to the original query. It is thus possible to efficiently search hundreds of millions of images, retrieving those that are not only textually related but also visually relevant. We demonstrated our approach on an application allowing users to search information on the web by matching images.

Typesetting math: 81%

In the future, we would like to explore more sophisticated image matching techniques and adapt them to the landmark image domain. In addition, we plan to carry out an extensive quantitative evaluation of our approach and an overall usability user study of the system.

Authors

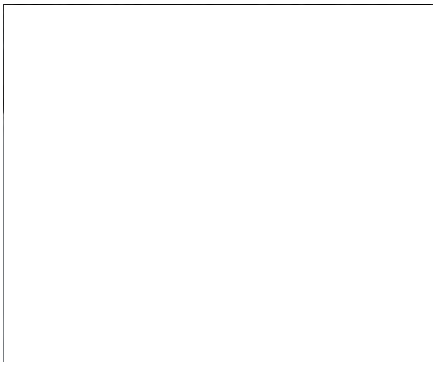
Figures

References

Citations

Keywords

Metrics



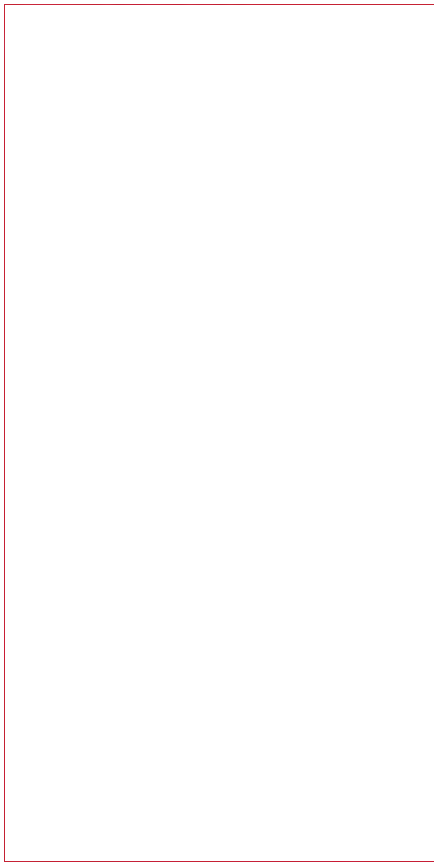
**More Like This**

An eigen-index technique for content-based retrieval of satellite image databases
IEEE International Geoscience and Remote Sensing Symposium
Published: 2002

A effective clustering indexing based on EMD for Web image retrieval
2008 First IEEE International Conference on Ubi-Media Computing
Published: 2008

**Show More**

Typesetting math: 81%



IEEE Personal Account

CHANGE USERNAME/PASSWORD

Purchase Details

PAYMENT OPTIONS

VIEW PURCHASED DOCUMENTS

Profile Information

COMMUNICATIONS PREFERENCES

PROFESSION AND EDUCATION

TECHNICAL INTERESTS

Need Help?

US & CANADA: +1 800 678 4333

WORLDWIDE: +1 732 981 0060

CONTACT & SUPPORT

Follow

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | IEEE Ethics Reporting | Sitemap | IEEE Privacy Policy

A not-for-profit organization, IEEE is the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity.

© Copyright 2024 IEEE - All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies.

IEEE Account

» Change Username/Password

» Update Address

Purchase Details

» Payment Options

» Order History

» View Purchased Documents

Profile Information

» Communications Preferences

» Profession and Education

Typesetting math: 81%

» Technical Interests

**Need Help?**

» **US & Canada:** +1 800 678 4333

» **Worldwide:** +1 732 981 0060

» Contact & Support

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | Sitemap | Privacy & Opting Out of Cookies

A not-for-profit organization, IEEE is the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity.

© Copyright 2024 IEEE - All rights reserved. Use of this web site signifies your agreement to the terms and conditions.

Typesetting math: 81%

**Appendix YEH02**

# Searching the Web with Mobile Images for Location Recognition

Tom Yeh, Konrad Tollmar, Trevor Darrell
MIT CSAIL
Cambridge, MA 02319

## Abstract

*In this paper, we describe an approach to recognizing location from mobile devices using image-based web search. We demonstrate the usefulness of common image search metrics applied on images captured with a camera-equipped mobile device to find matching images on the World Wide Web or other general-purpose databases. Searching the entire web can be computationally overwhelming, so we devise a hybrid image-and-keyword searching technique. First, image-search is performed over images and links to their source web pages in a database that indexes only a small fraction of the web. Then, relevant keywords on these web pages are automatically identified and submitted to an existing text-based search engine (e.g. Google) that indexes a much larger portion of the web. Finally, the resulting image set is filtered to retain images close to the original query. It is thus possible to efficiently search hundreds of millions of images that are not only textually related but also visually relevant. We demonstrate our approach on an application allowing users to browse web pages matching the image of a nearby location.*

## 1. Introduction

Content-based image retrieval (CBIR) has been an ongoing research topic for the past decade. One of the difficulties reported by users is that queries are hard to specify with the conventional input modalities; unlike keywords, a visual query may require the user to sketch or select parameter values for each low-level feature [10]. For recognizing real-world entities, searching with an actual image seems ideal. The advent of a new generation of camera-equipped phones and PDAs provides a unique platform for image-based web search.

We apply this paradigm to location recognition. Location recognition with radio or other signals (e.g., GPS, RFID tags, etc.) is possible, but an image-based approach can recognize locations distinct from a device's actual position. (e.g., "Tell me about that building over there.", while taking a snapshot of the building.) We have found that many landmarks and other prominent locations already have images on relevant web pages, so matching the mobile image to web documents can provide a reasonable location cue. We take images of an unknown location from a mobile camera, send them to a server for processing, and return a set of matching web pages.

Content-based search over the entire web is a computationally prohibitive proposition. Searching an image database on a scale similar to any commercial keyword-based image search engine, such as Google with 425 million images, remains a research challenge. We propose a hybrid approach that can provide content-based image retrieval by combining both image-based and text-based search. We leverage the fact that there will be redundant landmark images on the web and that a search in a subset of the web will likely find a matching image.

Our approach is to first search a bootstrap set of images obtained by web-crawling a restricted domain. The domain is selected based on the expected application, for example tourism-related sites for a particular geographic location. To recover relevant pages across the full web, we exploit a keyword-based search followed by a content-based filtering step. Keywords are extracted from web pages with matching images in the bootstrap set. Instead of running CBIR over hundreds of millions of images, we only need to operate on a seed set of images and on the images returned from the keyword-based search (see Figure 1).

In this paper, we focus on one particular class of images—landmarks for recognizing location. The goal is to recognize the location by finding web pages with an image matching the image of the location. In the next section we will review relevant related work. Following that, we present our approach to keyword and image-based web search and detail the component modules of our method. We present experiments on a tourist location recognition task to evaluate the components of our approach. We conclude with summary of our contribution.

## 2. Previous Work

Most commercially successful image search engines are text-based. Corbis features a private database of millions of high-quality photographs and artwork that are manually tagged with keywords and organized into categories. Google has indexed more than 425 million web pages and

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)
1063-6919/04 $20.00 © 2004 **IEEE**

IEEE COMPUTER SOCIETY

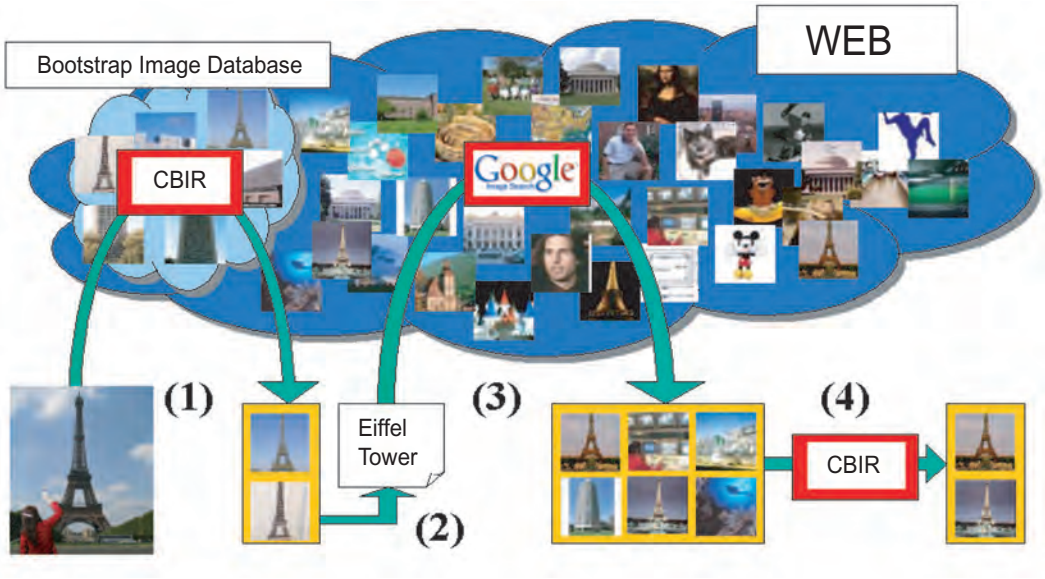



Figure 1: Image-and-keyword hybrid search. In step 1, the user takes an image with a camera phone, which is used as a query to find similar images from a small image database using CBIR techniques. In step 2, keywords are automatically extracted. In step 3, extracted keywords are sent to Google to find textually related images. In step 4, content-based image matching techniques are applied to identify visually relevant images.

inferred their content in the form of keywords by analyzing the text on the page adjacent to the image, the image caption, and other text features. In both cases, the image search engine searches for images based on text keywords. Since the visual content of the image is ignored, images that are visually unrelated can be returned in the search result. However, this approach has the advantage of text search—semantically intuitive, fast, and comprehensive.

Searching for images from images is often called content-based image retrieval and has been an ongoing research topic for many years [10]. Although the algorithms for analyzing image content in general run significantly slower than those for analyzing text, the search result can exhibit stronger visual relevance; we can have greater confidence as to whether the images retrieved fit the intended query. One of the first such systems was IBM's Query-By-Image-Content (QBIC) system [3]. It supported search by example images, user-drawn pictures, or selected color and texture patterns, and was applied mainly to custom, special-purpose image databases. In contrast, the Webseek system [11] searched generically on the World Wide Web for images. This system incorporated both keyword-based and content-based techniques; the keyword search returned a set of images among which users could further search by color histogram similarity with relevance feedback. The Diogenes system used a similar bi-modal approach for searching images of faces on the web [2]. A face detector operated as a screening process to filter out nonface images from the set of initial images returned by a keyword-based query. Then, a face recognition module processed the remaining images to identify particular faces.

Describing what images to search for can be tricky. QBIC lets users choose the relative weights of each feature in determining image similarity. Lew et. al proposed a system that allows users to sketch the outline of the image [6]. VisualSeek features an interface to paint a color template for template-based image matching [11]. However, these systems failed to provide an intuitive interface. The correspondence between low-level parameter values and high-level visual concept is not so obvious to ordinary users. Drawing can require too much effort and artistic skill. One solution is the search-by-example approach. For example, ImageRover presents the user with a set of starting images as examples and uses a "relevance feedback" framework to refine the search iteratively based on what the user says about the relevance of each image [9].

Due to the complexity typically involved in CBIR algorithms, there is a tradeoff between interactivity and comprehensiveness. Many existing CBIR systems use simple nearest-neighbor techniques to find matching images, which cannot scale to a large and comprehensive image database and still perform fast enough to be interactive. Hierarchical indexing with Self-Organizing Maps [5] and related techniques is possible, but the speed improvement is very limited. So far, practical CBIR systems operate on a scale much smaller than their text-based counterparts. We

were not aware of any CBIR system that had gone beyond the million images mark in scale [6, 11, 9, 2, 3].

Our approach to image-based location-recognition is closely related to efforts in the wearable-computing and robotics literature on navigation and scene recognition. The wearable museum guiding system built by Starner and colleagues uses a head-mounted camera to record and analyze the visitor's visual environment [12]. Computer vision techniques based on oriented edge histograms were applied to recognize objects in the field of view. Based on the objects seen, the system estimates the location in the museum and displays relevant information. The focus of this system was on recalling prior knowledge of locations—which item is exhibited where—rather than finding new information about locations. Torralba et al. generalized this type of technique and combined it with a probabilistic model of temporal dynamics, so that information from a video sequence from a wearable device could be used for mapping and navigation tasks [13]. In these robotics and wearable computing systems, recognition was only possible in places where the system had physically been before. In our system location-relevant information can be found from a single image of a novel place.

# 3. Image-based Web Search

Web authors tend to include semantically related text and images on web pages. There exists a high correlation between semantic relevancy and spatial proximity that can be exploited on the web; pieces of knowledge close together in cyberspace tend to be also mutually relevant in meaning [1]. To find information about a well-known landmark, we can search for web pages with images matching the image of the current location, and analyze the surrounding text. Using this approach, the location-recognition problem can be cast as a CBIR problem—if methods can be found to match mobile images to the web despite time, pose, and weather variation, they can serve as a useful tool for mobile web search (and in the particular application we consider here, location-based computing.)

With a camera phone, a user can take an picture of a prominent landmark around the current location. This picture can be used as a query to find images similar to this landmark on the web using content-based image matching techniques. Relevant keywords can be found in the surrounding text and used directly as a location context cue, or used for further interactive browsing to find relevant information resources. A very attractive property of this method is that it requires no special-purpose communications infrastructure or prior location database, and that from a user-interface perspective users can specify nearby locations with a very natural interface action—taking a picture of the intended place.

For a pure CBIR system to match against the millions of images on the web in real-time is currently computationally infeasible. However, using a hybrid keyword-and-image query system, we effectively implement CBIR over the entire 425 million images without having to apply a content-based metric on every single image. Such a hybrid design benefits from the power of both keyword-based search (speed and comprehensiveness) and image-based search (visual relevancy).

We leverage an existing keyword-based image search engine, Google, which has indexed more than 425 million images. We extract keywords from web pages found in a content-based search in a bootstrap database, and use these keywords on Google to search its larger database of images for images we want. Recall that one shortcoming of keyword-based search is the existence of visually unrelated images in the result set. To overcome this, we apply a filtering step, where we run CBIR on this small set of images to identify visually related images. In this way we retrieve images that are not only visually relevant but also textually related.

## 3.1. Image Matching Metrics

Having the right feature set and image representation is very crucial for building a successful CBIR system. The performance of general object matching in CBIR systems is far from perfect—image segmentation and viewpoint variation are significant problems. Fortunately, finding images of landmarks requires analysis over the entire image, making general image segmentation unnecessary. Also, users ask about a location most likely because they are physically there and there are a much smaller number of physically common viewpoints of prominent landmarks than in the entire viewsphere of a common object. Consequently, the simplicity of recognizing location images naturally makes it an attractive test-bed to test the idea of an image-and-text hybrid search approach.

We have experimented with two image-matching metrics on the task of matching mobile location images to images on the World Wide Web. The first metric is based on the *energy spectrum*, the squared magnitude of the windowed Fourier transform of an image. Given an image $i$, the discrete Fourier transform can computed as

$$I(f_x, f_y) = \sum_{x,y=0}^{N-1} i(x,y)h(x,y)e^{-j2\pi(f_x x + f_y y))}$$

where $h(x,y)$ is a circular hanning window to reduce the boundary effect. The amplitude component of the above term represents the spatial frequency spread everywhere in the image. It embeds unlocalized information about the image structure such as orientation, smoothness, length, and width of the contours that compose the entire scene in the

image. This type of representation was demonstrated by [13] to be invariant to object arrangement and object identity.

The second metric is based on wavelet decompositions. We compute the local texture features of an image $I$ by convolving it with steerable filters [4] over 2 different scales on the intensity (grayscale) image to obtain a representation as follows:

$$\lambda = G_{\theta_i}(S_j(I))$$

where $G$ is the steerable filter for 6 different $\theta$s, evenly spaced between 0 and $2\pi$, and $S$ is the scaling operator. This can produce $6 \times 2 = 12$ subbands. Since this gives us only the local representation of the image, we take the mean values of the magnitude of the local features averaged over large windows to capture global image properties

$$m(x) = \sum_{x'} |\lambda(i)| \cdot w(x' - x)$$

where $w$ is the averaging window. The final representation is downsampled to have a spatial resolution of 4 by 4 pixels resulting in a dimensionality of $4 \times 4 \times 12 = 96$.

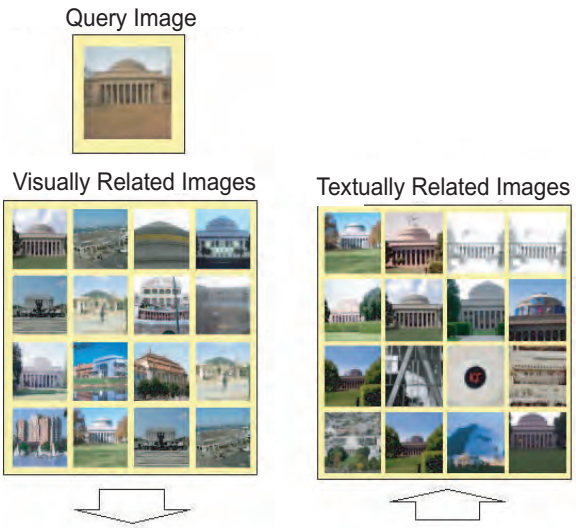
Given a mobile image of some landmark, similar images can be retrieved by finding the $k$ nearest neighbors in the database using either of the two metrics (e.g. see Figure 2, where $k = 16$). However, the high dimensionality ($d$) of the feature involved in the metric can be problematic. To reduce the dimensionality, we computed principal components over a large number of landmark images on the web. Then, each feature vector can be projected onto the first 32 (a number picked empirically) principal components which form the final feature vector for the images.

### 3.2. Extracting Useful Textual Information

After similar landmark images are found, we need to analyze the text of the web documents containing these images with the intention to uncover any useful information about this particular landmark. We devise a simple frequency-based approach to identifying named-entities (e.g. name, location, telephone number) by computing the *term frequency inverse document frequency* metric $\mu$ for each bigram and trigram found in the document:

$$\mu(w) = \frac{df(w)}{tf(w)}$$

where $tf(w)$ is the global frequency of $w$ over a large corpus and $df(w)$ is the local frequency of $w$ within the document [8]. The underlying intuition is that people tend to find the sudden occurrence (high $df$) of uncommon things (low $tf$) interesting. A good estimate of the global word frequency can be the number of links Google indexes for this word. The top $n$ word combinations with the highest $f$ can be reported as the potential named-entities to the users.




| Unigram | $\mu$ | Bigram | $\mu$ |
|---|---|---|---|
| MIT | 128 | foundation relations | 3 |
| story | 59 | MIT dome | 2 |
| engineering | 33 | da lucha | 2 |
| kruckmeyer | 29 | view realvideo | 2 |
| boston | 28 | cancer research | 2 |

Figure 2: A typical search scenario: query image to visually similar images to named-entities to textually relevant Google images.

Figure 2 illustrates a typical search scenario. A user initiates a search request by taking an image of some prominent landmark (e.g. MIT Dome) at the current location. Sixteen visually similar images to the query image are found in the web image database. Five of them are correct (1,3,4,9, and 14). The content of the source HTML documents of these 5 images are analyzed and potential named-entities are found and ranked. These named-entities can be submitted as keywords to a text-based search engine (e.g. Google) to find more images. However, it can not be guaranteed that these images are visually similar to the questioned landmark.

### 3.3. Content-filtered Keyword Search

The name-entities extracted from the web pages can provide the users with not only the basic notion about this location but also the possibility to acquire further information from a powerful text-based search engine (e.g. Google).

One possibility is to search for more web pages that might lack any visual content but nevertheless indirectly share similarity with the query image through text. We can describe such similarity inference notationally:

$$I_q \underset{image}{\longleftrightarrow} I_v, P_v \underset{text}{\longleftrightarrow} P_t \Longrightarrow I_q \longleftrightarrow P_t$$

where $I_q$ denotes the query image, $I_v$ denotes the found similar image, $P_v$ denotes the web page containing $I_v$, $P_v$ denotes a web page containing similar keywords as $P_t$, and $\leftrightarrow$ denotes a similarity relation based on either text or image. Here, $P_t$ would not have been found otherwise if either of the pure image-based search or pure text-based search were employed.

Another possibility is to use the keywords to search for images $I_t$. The similarity inference can be denoted as:

$$I_q \underset{image}{\longleftrightarrow} I_v, P_v \underset{text}{\longleftrightarrow} I_t \Longrightarrow I_q \longleftrightarrow I_t.$$

While $I_q$ and $I_t$ are textually related, to guarantee visual similarity, a content-based filtering step needs to be applied to keep keep only those images visually close to the query image. As shown in Figure 4, a search on the query image $I_q$ (a 15 floor building) yields the keywords "MIT Green Building". These keywords are submitted to Google to retrieve 16 such $I_t$s that are textually relevant but not quite visually similar. Sorting each $I_t$ according to its distance to $I_q$, we are able to boost the ranks of the visually similar images to the top.

Moreover, there might exist images visually distant but conceptually close to the query image; they can help us learn more about this location. We can use a bottom-up, opportunistic clustering technique that iteratively merges data points to uncover visually coherent groups of images. Initially, each image is considered as the center in its own cluster. At each iteration, two closest clusters, in terms of the same distance metric used in other parts of the system, are merged. The process will continue until the distance between every cluster pair has exceeded a certain threshold. If a cluster is reasonably large, it means the images in this group represent some potentially significant common concept. In the example shown in Figure 4, four images are removed from the result set because they can not be clustered with any other image.

# 4. Experiments

A series of experiments was conducted to evaluate the effectiveness of our algorithms for location recognition by matching mobile images to a web database. We chose the task of recognizing landmark images on a university campus. We ran a web crawler on $mit.edu$ to collect images and their source URL's to build an image database as the bootstrap set. Images too small to contain interesting location information, less than 100 by 100, were discarded. The resulting database contained roughly 33,000 images. Since dealing with images of various dimensions and sizes can be tricky, we resized each image to 100 by 100 for uniformity.

To find similar landmark images, it would not be useful to search images that do not contain any landmark (e.g. faces, animals, or logos). Thus, an image classifier was used to classify these 33,000 images as landmark or nonlandmark. The nonlandmark images were then removed from the database to reduce the search space to 2000+ images. The image classifier was trained using a method similar to [7] that was demonstrated to be successful in classifying indoor-outdoor images by examining color and texture characteristics.

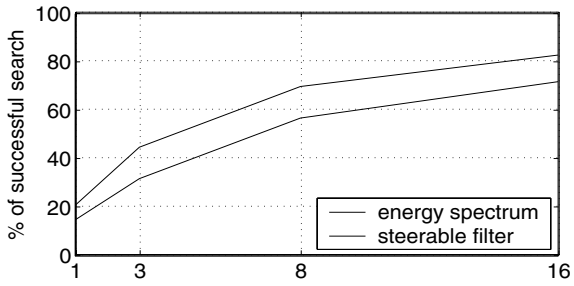



Figure 3: The performance of the matching metrics with k nearest neighbors.

Since we were concerned with location recognition in a mobile setting, the test queries consisted of landmark images acquired with a camera-equipped mobile device. We selected 5 prominent landmarks and asked volunteers to take 20 images of each of these landmarks using a Nokia 3650 camera phone. A total of 100 query images were obtained for 100 test runs.

The matching metrics discussed in section 3.1 were used in our experiments. For each query image, the $k$ nearest images were retrieved from the database to form the search result. The same volunteers were asked to evaluate the search result. A success was defined as the occurrence of at least one correct landmark image among the first $k$ nearest images. The findings are summarized in Figure 3, which plots the percentage of test cases that were successful over different $k$.

At $k = 16$, the likelihood of finding correct images in the result set is reasonably high. This indicates that even with these standard feature sets it is possible to find similar location images for recognition tasks.

# 5. Conclusions and Future Works

In this paper, we explored the potential of web search with mobile images using both images and text. We have shown it is possible to conduct fast and comprehensive CBIR over hundreds of millions of images using a text-based search engine on keywords generated from an initial search. We presented an approach to recognizing location from mobile devices using image-based web search, and showed that common image search metrics can match images captured with a camera-equipped mobile device to images found on the

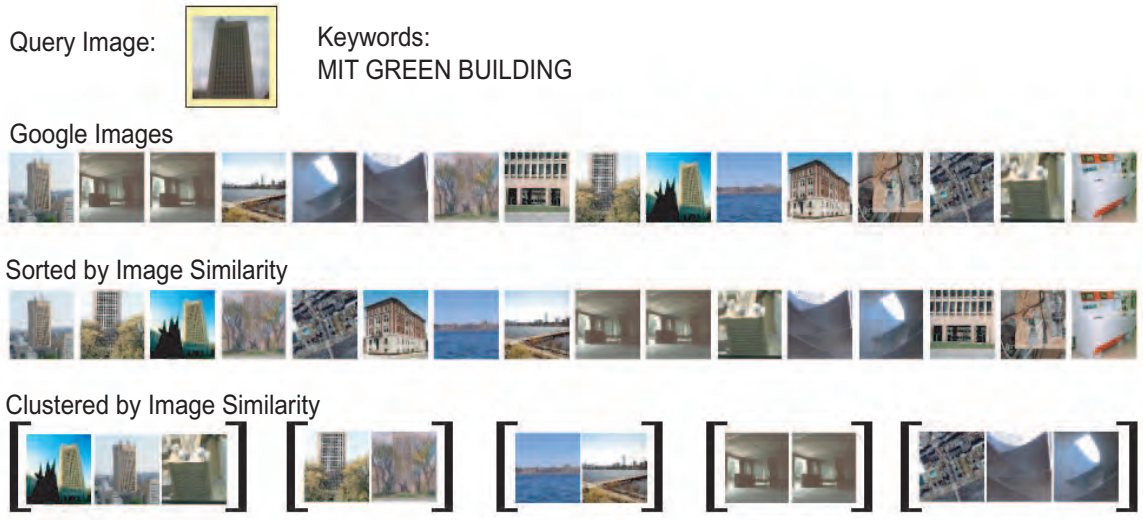



Figure 4: Content filtered keyword search for images. The keywords found for the query image are submitted to Google and a set of textually released images are found. These images can be further filtered according to their visual similarity to the query image either by sorting or clustering.

World Wide Web or other general-purpose databases.

A hybrid image-and-keyword searching technique was developed that first performed an image-based search over images and links to their source web pages in a bootstrap database that indexes only a small fraction of the web and a simple frequency-based procedure was employed to extract relevant keywords from these web pages. These keywords can be submitted to an existing text-based search engine (e.g. Google) that indexes a much larger portion of the web. The resulting image set is then filtered to retain images close to the original query. It is thus possible to efficiently search hundreds of millions of images, retrieving those that are not only textually related but also visually relevant. We demonstrated our approach on an application allowing users to search information on the web by matching images.

In the future, we would like to explore more sophisticated image matching techniques and adapt them to the landmark image domain. In addition, we plan to carry out an extensive quantitative evaluation of our approach and an overall usability user study of the system.

## References

[1] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the web. *ACM Transactions on Internet Technology*, 1(1):2–43, 7 2001.

[2] Y.A. Aslandogan and C. Yu. Multiple evidence combination in image retrieval: Diogenes searches for people on the web. In *Proceedings of ACM SIGIR 2000*, pages 88–95, 2000.

[3] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, and B. Dom. Query by image and video content: the QBIC system. *IEEE Computer*, 28(9):23–32, 1995.

[4] W.T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.

[5] J. Laaksonen, M. Koskela, S. Laakso, and E. Oja. PicSOM: Content-based image retireval with self-organizing maps. *Pattern Recognition Letters*, 21(13-14):1199–1207.

[6] M. Lew, K. Lempinen, and N. Huijsmans. Webcrawling using sketches. pages 77–84, 1997.

[7] Szummer M. and R. W. Picard. Indoor-outdoor image classification. In *IEEE Intern. Workshop on Content-based Access of Image and Video Databases*, pages 42–51, 1998.

[8] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[9] Stan Sclaroff, Leonid Taycher, and Marco LaCascia. ImageRover: A content-based image browser for the world wide web. Technical Report 1997-005, 24, 1997.

[10] Worring M. Santini S. Gupta A. Smeulder, A. and R. Jain. Cbir at the end of the early years, Dec 2000.

[11] John R. Smith and Shih-Fu Chang. Image and video search engine for the World Wide Web. In *Proc. of SPIE)*, pages 84–95, 1997.

[12] T. Starner, B. Schiele, and A. Pentland. Visual contextual awareness in wearable computing. In *Proc. of Visual Contextual Awareness in Wearable Computing*, pages 50–57, 1998.

[13] A. Torralba, K.P. Murphy, W.T. Freeman, and M.A. Rubin. Context-based vision system for place and object recognition. In *Proc. of ICCV*, pages 273–280, 2003.