

Digital
Multimedia
Standards
Series

MPEG VIDEO COMPRESSION STANDARD

Joan L. Mitchell,
William B. Pennebaker,
Chad E. Fogg, and
Didier J. LeGall

MPEG VIDEO COMPRESSION STANDARD

For all those interested in high definition television, multimedia, and image compression, this unique reference will be an essential tool. It provides the first comprehensive introduction to this field, incorporating material ranging from basic concerns of newcomers to the field through sophisticated reviews of cutting edge technical issues.

Written by acknowledged experts in the field, **MPEG Video Compression Standard** offers important benefits to readers including:

- Detailed information on MPEG modes of operation, signaling conventions, and structure of MPEG compressed data
- Each section of the book is labeled by level of technical difficulty, allowing less technical readers to skip higher level sections and still gain a broad understanding of the subject while guiding advanced readers to the in-depth coverage they require

With its comprehensive coverage of MPEG video compression, this important book meets the needs of those working to develop the standard as well as those who use MPEG in their work. Electrical engineers, multimedia producers, computer scientists, as well as all those interested in this fast growing field will find **MPEG Video Compression Standard** essential in their work.

ABOUT THE AUTHORS

Joan L. Mitchell is a Research Staff Member at the IBM Thomas J. Watson Research Center.

William B. Pennebaker is a consultant and former Research Staff Member, IBM Thomas J. Watson Research Center.

Chad E. Fogg, Cascade Design Automation

Didier J. LeGall is Vice President of Research and Development, C-cube Microsystems.



Chapman & Hall
115 Fifth Avenue, New York, NY 10003
2-6 Boundary Row, London SE1 8HN
Printed in U.S.A.

JOIN US ON THE INTERNET WWW: <http://www.thomson.com>

Visit Chapman & Hall's Internet Resource Center for information on our new publications, links to useful sites on the World Wide Web and an opportunity to join our e-mail mailing list.

Point your browser to: <http://www.chaphall.com/chaphall.html> or <http://www.chaphall.com/chaphall/electeng.html> for Electrical Engineering

A service of **ITP**[®]

ISBN 0-412-08771-5



9 780412 087714

Digital
Multimedia
Standards
Series

MPEG VIDEO COMPRESSION STANDARD

Joan L. Mitchell,
William B. Pennebaker,
Chad E. Fogg, and
Didier J. LeGall



CHAPMAN & HALL



INTERNATIONAL THOMSON PUBLISHING

New York • Albany • Bonn • Boston • Cincinnati • Detroit • London • Madrid • Melbourne
Mexico City • Pacific Grove • Paris • San Francisco • Singapore • Tokyo • Toronto • Washington

Cover design: Curtis Tow Graphics

Copyright © 1996 by Chapman & Hall

Printed in the United States of America

Chapman & Hall
115 Fifth Avenue
New York, NY 10003

Chapman & Hall
2-6 Boundary Row
London SE1 8HN
England

Thomas Nelson Australia
102 Dodds Street
South Melbourne, 3205
Victoria, Australia

Chapman & Hall GmbH
Postfach 100 263
D-69442 Weinheim
Germany

International Thomson Editores
Campos Eliseos 385, Piso 7
Col. Polanco
11560 Mexico D.F.
Mexico

International Thomson Publishing-Japan
Hirakawacho-cho Kyowa Building, 3F
1-2-1 Hirakawacho-cho
Chiyoda-ku, 102 Tokyo
Japan

International Thomson Publishing Asia
221 Henderson Road #05-10
Henderson Building
Singapore 0315

All rights reserved. No part of this book covered by the copyright hereon may be reproduced or used in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or information storage and retrieval systems—without the written permission of the publisher.

2 3 4 5 6 7 8 9 10 XXX 01 00 99 98 97

Library of Congress Cataloging-in-Publication Data

MPEG video : compression standard / Joan L. Mitchell . . . [et al.].
p. cm.

Includes bibliographical references and index.
ISBN 0-412-08771-5 (alk. paper)

1. Digital video. 2. Video compression -- Standards. 3. Sound -
- Recording and reproducing -- Digital techniques -- Standards.
4. Coding theory. I. Mitchell, Joan L.
TK6680.M64 1996
621.388--dc20

96-31124
CIP

British Library Cataloguing in Publication Data available

To order this or any other Chapman & Hall book, please contact **International Thomson Publishing, 7625 Empire Drive, Florence, KY 41042**. Phone: (606) 525-6600 or 1-800-842-3636.
Fax: (606) 525-7778. e-mail: order@chaphall.com.

For a complete listing of Chapman & Hall titles, send your request to **Chapman & Hall, Dept. BC, 115 Fifth Avenue, New York, NY 10003**.

Introduction

The subject of this book is the international standard for moving picture video compression, IS 11172-2, widely known as MPEG-1 video. The name MPEG is an acronym for Moving Picture Experts Group, a group formed under the auspices of the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC). Originally an ad hoc subcommittee of ISO/IEC JTC1/SC2/WG8 where the MPEG work started, MPEG was later given formal status within ISO/IEC and is now the working group ISO/IEC JTC1/SC29/WG11.

Although IS 11172 is officially entitled “Information technology — Coding of moving pictures and associated audio for digital storage media up to about 1,5 Mbit/s —,”¹ as indicated by the title, MPEG-1 is concerned with coding of digital audio and digital video. It is also concerned with synchronization of audio and video bitstreams, including multiple interleaved video sequences. These three topics are covered in three separate parts of the standard: Part 1 deals with systems aspects, Part 2 with video compression, and Part 3 with audio compression.²

In this book the emphasis will be almost entirely on the compression of picture sequences, i.e., on IS 11172 Part 2: *Video*. While brief overviews of the audio and systems aspects of MPEG-1 are provided, it is our expectation that future volumes in this series will do more justice to these important topics.

If the current standard is known as MPEG-1, what does the “1” signify? It turns out that there are three different MPEGs: MPEG-1, MPEG-2, and MPEG-4,³ and data rate and applications are distinguishing factors

¹Note the European convention of a comma for the decimal point. We will follow the U.S. convention and use a period, unless we are quoting from the standard.

²Part 4 covers compliance testing, and Part 5 is a technical report on software implementation.

³There was once an MPEG-3, but it was merged with MPEG-2.

between them. MPEG-1 is intended for intermediate data rates on the order of 1.5 Mbit/s, MPEG-2 is intended for higher data rates (10 Mbit/s or more), and MPEG-4 is intended for very low data rates (about 64 Kbit/s or less). MPEG-2 and MPEG-4 have potential uses in telecommunications, and for this reason, a third major international standards organization, the International Telegraph and Telephone Consultative Committee (CCITT), is participating in their development.⁴ At this point in time, MPEG-1 and MPEG-2 video are complete (some later parts of MPEG-2 are still being written), and MPEG-4 is still in the algorithm development stage.

One version of MPEG-2 video, the *main profile*, is a relatively straightforward extension of MPEG-1. Therefore, while this book is concerned primarily with MPEG-1, an overview of the main profile of MPEG-2 will also be given. We note, however, that there is a lot more to MPEG-2, including four other profiles that provide major extensions in flexibility and function. Another volume in this series provides a far more comprehensive treatment of MPEG-2 [HPN97].

1.1 Why compress? ○

A sequence of pictures with accompanying sound track can occupy a vast amount of storage space when represented in digital form. For example, suppose the pictures in a sequence are digitized as discrete grids or arrays with 360 pels (picture elements) per raster line and 288 lines/picture, a resolution that is fairly typical for MPEG-1. Assuming the picture sequence is in color, a three-color separation (equivalent to the three primaries, red, green, and blue) can be used for each picture. If each color component in the separation is sampled at a 360x288 resolution with 8-bit precision, each picture occupies approximately 311 Kbytes. If the moving pictures are sent uncompressed at 24 pictures/second, the raw data rate for the sequence is about 60 Mbit/s, and a one-minute video clip occupies 448 Mbytes.

For the audio track the data rates are not quite as formidable. If the sound track is in stereo and each of the two channels is sampled with 16-bit precision at a 44 KHz sampling rate, the data rate is about 1.4 Mbit/s.

The problem MPEG-1 addresses is how to store a moving picture sequence, including the audio information, in a compressed format suitable for a digital storage medium such as CD-ROM (compact disc - read only memory). The magnitude of the problem surfaces immediately when one reflects that this is exactly the same digital storage technology that is used for high-quality compact disc (CD) sound reproduction. A high-quality stereo sound track *plus* the video information must be stored in about the same

⁴The name of the CCITT was recently changed to International Telecommunications Union-Telecommunication Sector (ITU-T).

space that is normally used for the audio alone. Further, this must be done in a way that permits playback with precise timing control and synchronization of audio and video bitstreams.

1.2 Why standardize video compression? ○

When the MPEG standardization effort was initiated in 1988, several different industries were converging on digital video technology. The computer industry was looking to expand beyond its traditional text and graphics capabilities into interactive multimedia with audio, video, and still images. The consumer electronics industry saw digital video as a means for improving the capabilities and interactivity of video games and entertainment media. Compact discs were already being used for digital storage, and the storage capacities were sufficient for compressed digital video. The telecommunications industry was taking advantage of the maturing compression technology to standardize teleconferencing, and the CCITT Recommendation H.261 digital video standard for teleconferencing was already in development. The debate about broadcasting a digital versus a higher bandwidth analog signal for high-definition television (HDTV) was about to start and direct broadcast of digital video from satellites was also being considered.

Cable television companies were exploring the use of digital video for satellite uplinks and downlinks, and they were also considering delivery of digital video to the home. Together with the microprocessors that would be used for control of home converters, direct digital links into the home raised the possibility of new function through the cable networks. Other links into the home such as Videotex also had potential uses for moving sequences.

With the sharing of technology among these diverse industrial interests, standardization was a natural development. If standardization of digital video could be achieved, the potentially large sales volumes in the consumer market would lower the cost of VLSI chips for all of the industries. Thus, the cost reduction implicit in having a common digital video technology for all has been a key driving force for standardization. The convergence of these interests and the resulting standardization is lowering the barriers that are usually present to the deployment of new technology, and therefore reducing the risks.⁵

The MPEG-1 algorithm development was in its final stages, but had not completed the standards balloting process, when MPEG-2 was started.

⁵Not everyone would agree with this thesis. A number of proprietary teleconferencing systems have been marketed over the past decade, and a video compression technique known as Indio is finding use in multimedia. However, some of the individuals and corporations that were involved in these alternatives have also actively contributed to MPEG, and it remains to be seen how these other techniques will fare in relation to MPEG.

The major focus of this second project was higher quality and bandwidth applications, but there were some secondary issues such as the increased flexibility and multiple resolutions needed by some applications. Although the door was open to totally different technologies, it turned out that the best proposals for MPEG-2 were based on enhancements of MPEG-1. In the authors' opinion the rapid acceptance of MPEG-2 is in no small way due to the fact that it is a natural extension of MPEG-1. The similarities between MPEG-1 and the main profile of MPEG-2 also made it relatively easy for us to include that aspect of MPEG-2 in this book.

1.3 Vocabulary ○

A number of basic terms will be used throughout this book, and the following few paragraphs of definitions may be of help to the non-experts among our readers. More extensive discussions of these terms are found in [PM93] and [RJ92].

1.3.1 Sequences, pictures, and samples ○

An MPEG video sequence is made up of individual pictures occurring (usually) at fixed time increments, as shown in Figure 1.1. Because the pictures are in color, each picture must have three *components*. Color is expressed in terms of a *luminance* component and two *chrominance* components. The luminance provides a monochrome picture such as in Figure 1.4, whereas the two chrominance components express the equivalent of color hue and saturation in the picture. Although these components are not the familiar red, green, and blue (RGB) primaries, they are a mathematically equivalent color representation that can be more efficiently compressed. They are readily converted to RGB by relationships that will be described in Chapter 4.

Each component of a picture is made up of a two-dimensional (2-D) grid or array of samples. Each horizontal line of samples in the 2-D grid is called a raster line, and each sample in a *raster line* is a digital representation of the intensity of the component at that point on the raster line. Note, however, that the luminance and chrominance components do not necessarily have the same sampling grid. Because the eye does not resolve rapid spatial changes in chrominance as readily as changes in luminance, the chrominance components are typically sampled at a lower spatial resolution.

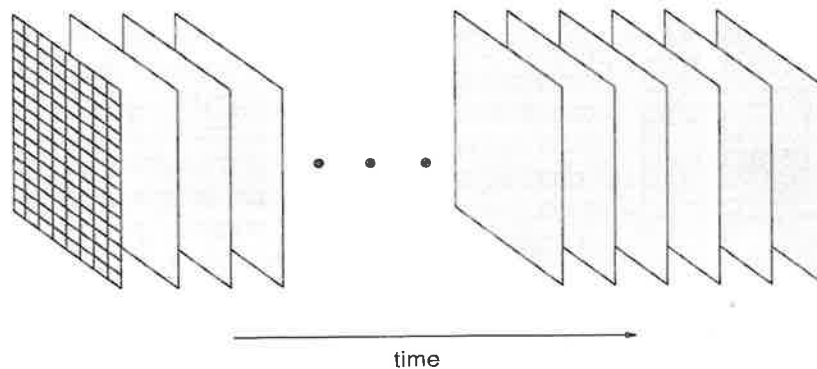


Figure 1.1: Illustration of an MPEG video sequence.

1.3.2 Frames and fields ○

In broadcast analog video standards such as NTSC, PAL, or SECAM⁶ video sequences are temporally subdivided into frames and raster lines in much the same manner as in Figure 1.1. However, the signal within each raster line is analog rather than digital. Further, as shown in Figure 1.2, each frame is divided into two *interlaced* fields. Each field has half the raster lines of the full frame and the fields are interleaved such that alternate raster lines in the frame belong to alternate fields. The frame repetition rate is therefore half the field rate.⁷

The relationship between analog video frames and fields and MPEG pictures is undefined for MPEG-1 and the assumption is made that appropriate 2-D pictures will be extracted from this field/frame format. The extensions in MPEG-2 allow interlaced pictures, coding them either as individual fields or full frames. Artifacts due to blanking are usually cropped before presenting the pictures to an MPEG encoder. The typical TV uses overscan to keep the blanking intervals hidden, and these areas are therefore normally not visible. An MPEG decoder must in some clever way recreate these cropped edge areas, and a poorly adjusted TV set might therefore exhibit some curious artifacts near picture edges.

⁶NTSC, PAL and SECAM are three commonly-used analog video formats. They are acronyms for National Television System Committee, phase alternating line, and sequentiel couleur a memoire, respectively.

⁷A vertical blanking interval separates the active display of each field, and this blanking interval typically inhibits the video signal for half of the top and bottom lines of field 1.

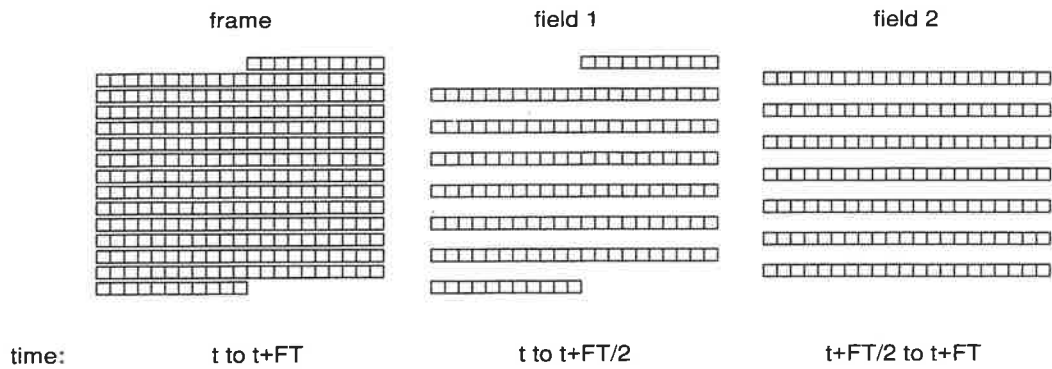


Figure 1.2: Analog video frame and field structure. FT is the time interval for a frame.

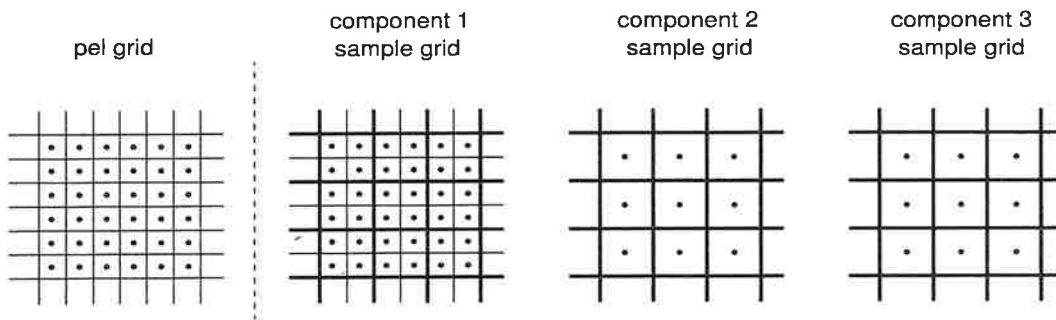


Figure 1.3: Component samples for MPEG-1 pels.

1.3.3 Pels and pixels ○

The component sample values at a particular point in a picture form a *pel* (picture element).⁸ If all three components use the same sampling grid, each pel has three samples, one from each component. However, as noted above, the eye is insensitive to rapid spatial variations in chrominance information. For this reason, MPEG-1 subsamples the components that supply the color information, using a chrominance grid that is a factor of 2 lower in resolution in each dimension. Thus, a pel is defined to be the color representation at the highest sampling resolution, but not all samples that make up the pel are at that resolution. MPEG-1 pels are sketched in Figure 1.3.

Note that when lower sampling resolution is used for some components

⁸ *Pixel* is also a commonly used abbreviation for picture element, but it is rarely used in the MPEG standard.

in a pel, the positioning of lower-resolution samples relative to the highest-resolution samples must be defined. In Figure 1.3 dots indicate the sample positions. The bold lines have been superimposed on the component 1 sample grid to illustrate the relative alignment of the component samples. In MPEG-1 the sample positioning is as in Figure 1.3; in MPEG-2 a different relative positioning is defined.

1.3.4 Compression vocabulary ○

The compression techniques used by MPEG are divided into two basic classes, *intra* and *nonintra*. Intra techniques compress a picture using information only from that picture; nonintra techniques also use information from one or two other pictures displaced in time.

The input to a data compressor is often termed the *source* data, whereas the output from a decompressor is called *reconstructed* data. Some compression techniques are designed such that reconstructed data and source data exactly match, and these techniques are termed *lossless*. With other techniques the reconstructed data are only a (hopefully) good approximation to the source data, and these techniques are termed *lossy*. The distinction between these two classes of compression techniques is extremely important, since purely lossless compression is normally far less efficient (in terms of bits/pel) than lossy compression. As will be discussed in Chapter 2, most data compression algorithms combine lossy and lossless elements. Very effective compression usually results when the lossy elements are designed to selectively ignore structure in the data that the eye cannot see (or the ear cannot hear).

1.3.5 Standards vocabulary ○

Standards documents separate material into two classes. *Normative* sections define the actual standard. *Informative* sections provide background material and descriptions, but do not in any way define the standard or any of the requirements and restrictions.

Standards documents also use a very particular vocabulary. Thus, the term *shall* has a particular meaning in “standardeze”, and is used to signify that the particular rule or requirement is normative and the implementor has no choice but to do exactly as the standard states. *May* indicates that the implementor is allowed some latitude. The rule suggested in the standard can be followed, but if the implementor so desires, he or she can follow an alternative or perhaps not implement that part of the standard at all. *Must* indicates that no alternative exists.⁹

⁹However, the MPEG standard does not always follow these rules rigorously. For example, in Clause 2.3 of the standard, there is the following statement: “In particular,

As is typical of standards documents, the MPEG standard usually states precisely how something “shall” be done, but is often missing the explanations and the rationales for why it is done in that particular way. One of the principal goals of this book is to provide these explanations.

1.3.6 MPEG vocabulary ○

As is true for any standard, MPEG uses a specialized vocabulary that is carefully defined in that particular standard. Most definitions are consistent with conventional technical usage. However, in some cases the term is more narrowly defined or may even be uniquely defined for that standard. These definitions are normative and take precedence over any conventional usage.

For example, consider the MPEG terms *access unit* and *presentation unit*. According to the definitions in subclause 2.1 of the MPEG-1 standard, these two terms apply to both video and audio bitstreams, but with different definitions. For audio, the presentation unit is a decoded audio access unit, and the audio access unit is the smallest independently decodable segment of an audio bitstream. For video, the presentation unit is a picture and the video access unit is the coded representation of that picture. Most pictures in a typical sequence are not independently decodable.

When reading the MPEG standard (or any standard, for that matter), care must be taken to identify this specialized vocabulary. Further, the definitions in force in each part of the standard must be clearly understood, or misinterpretation is possible. Unless we state otherwise, our usage in this book will be consistent with the definitions for MPEG video.

1.4 A quick look at MPEG compression ○

Although there are many facets to data compression and image coding, perhaps the key aspect of moving picture compression is the similarity between pictures in a sequence such as in Figure 1.4. The similarity is illustrated more clearly by taking differences between the pictures, as is done in Figure 1.5. In this difference sequence, pel differences of zero are neutral gray, positive differences are proportionally lighter, and negative differences are proportionally darker.

The most visible differences in Figure 1.5(a) occur where objects, such as the ball, are in motion. As a general rule, less activity (smaller differences)

it [subclause 2.4.3] defines a correct and error-free input bitstream. Actual decoders *must* [our emphasis] include a means to look for start codes in order to begin decoding correctly, and to identify errors, erasures or insertions while decoding. The methods to identify these situations, and the actions to be taken, are not standardized.” It is certainly possible to construct a decoder that does not have error recovery procedures, but it would be wiser to follow the MPEG committee requirement.

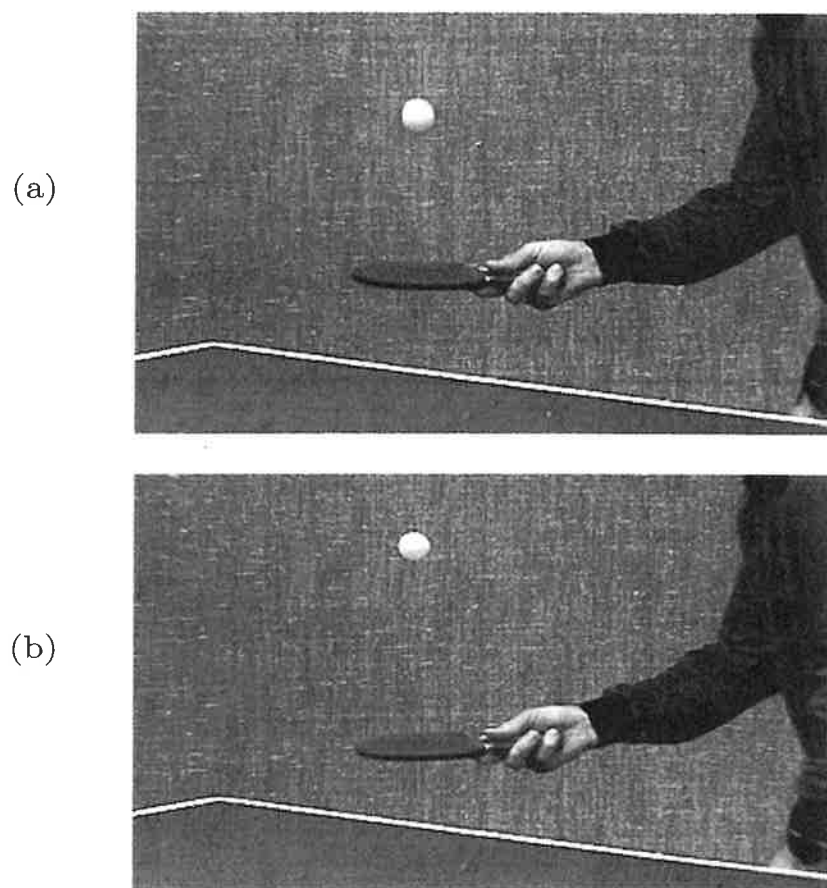


Figure 1.4: Pictures 30 (a) and 32 (b) of the tennis video sequence.

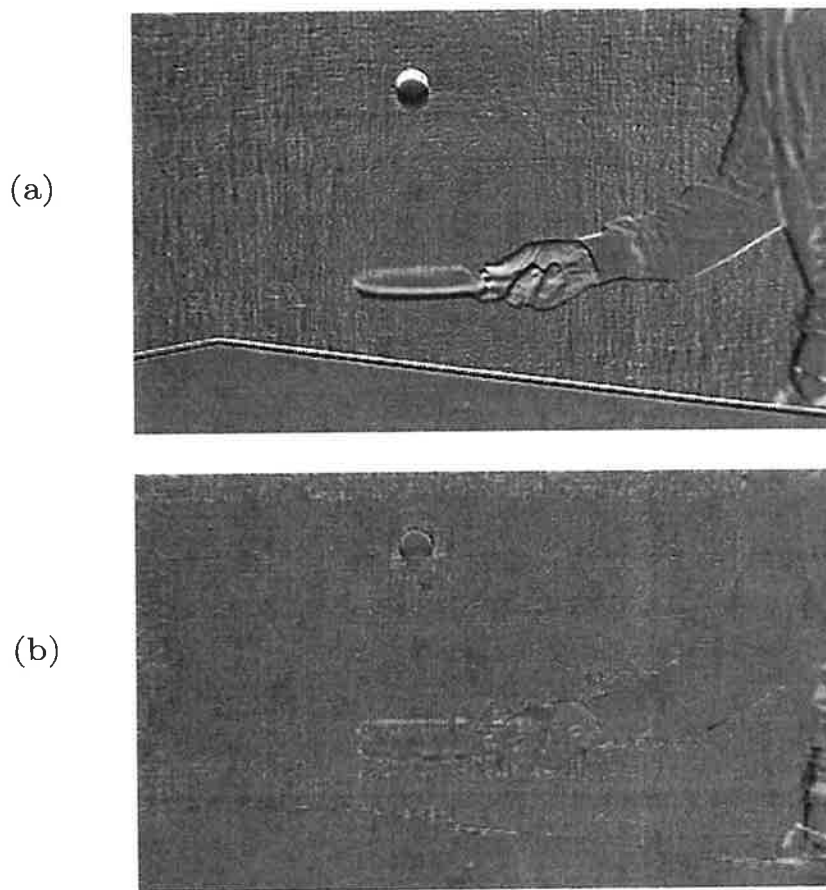


Figure 1.5: Picture differences: (a) simple differences. (b) motion-compensated differences.

in a picture leads to better compression. If there is no motion, as is the case in the center of the pictures, there is no difference, and the sequence can be coded as a single picture followed by a few bits for each following picture telling the decoder to simply repeat the picture. Not surprisingly, this is exactly what MPEG does in this situation.

Of course, coding a completely unchanging sequence is a very special case, and much more must be done to code a typical picture sequence. The sequence must start somewhere, and a technique is therefore needed for coding the first picture without reference to neighboring pictures. It turns out that this technique must also be used at regular intervals in a typical sequence. If pictures in a sequence are not occasionally coded independently of neighboring pictures, video editing becomes very difficult. Furthermore, because of bit errors during transmission or the small differences allowed between encoder and decoder image reconstructions, differences between the original source picture sequence and the sequence reconstructed by a decoder may build up to unacceptable levels.

The high compression needed by MPEG applications is achieved by coding most of the pictures as differences relative to neighboring pictures. MPEG accomplishes this in a variety of ways: Some parts of the picture — those where no significant changes occur — are simply copied. Other parts are imperfectly predicted from the adjacent picture and corrections must be coded. Still other parts may be best predicted by parts of the image that are displaced because of motion, and this requires the use of a technique called *motion compensation*. For example, in the peripheral areas of pictures in Figure 1.4, camera zoom caused a systematic displacement inward from (a) to (b). This is why the differences are more visible in the peripheral areas of Figure 1.5(b).

The differences are greatly reduced when the encoder searches picture (a) for a best match to (b), as is done in Figure 1.5(b). When parts of the picture are in motion, the best match for a given region between Figure 1.4(b) and (a) is shifted slightly because of the motion. If the position of (a) is shifted by that amount before taking the difference, Figure 1.5(b) results. Clearly, motion compensation eliminates much of the activity.

Some parts of a picture may have new content that does not exist in the reference picture — for example, areas uncovered when an object moves and new areas introduced by panning and zooming. This situation can be handled by coding these parts of the picture without reference to adjacent pictures. However, in many cases it can also be handled by predicting backward in time, assuming that an upcoming picture has already been coded and transmitted. An inviolable rule of data compression systems is that predictions can only be made from information available to the decoder.

A large asymmetry exists between MPEG encoders and MPEG decoders

in terms of complexity and computational requirements, and this is partly due to motion compensation. The encoder has the task of finding the motion displacements (*vectors*), whereas the decoder merely uses the values transmitted to it by the encoder. In other places as well, the encoder must traverse a difficult decision tree to determine what to send to the decoder; the decoder simply does what it is told to do.

1.5 Compact pseudocode notation in MPEG ○

Those readers who have already had some exposure to MPEG will know that the standard uses a compact pseudocode notation in the normative sections of the standard. This is a very precise notation that is closely related to the C language; however, it can sometimes be quite cryptic and difficult to understand. In the authors' experience, those who have mastered this notation are very enthusiastic about it. However, there is a steep learning curve — especially for readers who do not have a background in C programming — that makes it difficult for many to understand MPEG. One of the goals of this book is to provide an alternative to this compact notation that allows readers without a C language background to understand MPEG video.

In the final analysis, however, no one who really needs to understand a standard can avoid reading the actual document. For this reason, we will also use pseudocode in some of our explanations, but with one significant improvement — we have added comments to almost every statement. Comments are informative, and therefore would have been inappropriate in the normative sections of the MPEG standard that define the pseudocode. We feel, however, that they add enormously to the readability of the code.

1.6 MPEG applications ○

Although MPEG started out as an anticipatory standard, its acceptance has been so rapid that in many cases the applications have been ahead of the approval process. For example, even before the standard was approved, software, multimedia boards, and chips implementing MPEG-1 were already available. Games, digital recorders, CD-ROM movies, digital TV, and multimedia are just a few of the current applications. Indeed, the headline, “Crooks Crack Digital Codes of Satellite TV,” from the January 12, 1996 issue of the Wall Street Journal[TR96] hints at how financially important the digital satellite TV application has already become.

The committee made some effort to keep MPEG-1 stable and unchanged once technical consensus had been reached. This made possible these early implementations, but also prevented the standard from being modified to

meet the higher bandwidth needs addressed by MPEG-2. This was one of the key reasons for the separate development of MPEG-2.

The acceptance of MPEG-2 has been even more rapid. High-definition television (HDTV) standards in the United States, Europe, and Japan are expected to use MPEG-2, and it will also be used in direct broadcast of high-quality digital video at current resolutions. Indeed, in at least one case deployment of decoders for this latter application has been delayed until the MPEG-2 standard is in final form [Rob94].

1.7 Organization of the book ○

The more technical sections of this book assume a basic knowledge of image coding concepts. However, we have included overview sections at the start of each chapter that provide background for non-technical readers. The chapters of the book are organized as follows:

Chapter 2 discusses the full MPEG-1 system, providing an overview of all three parts of the Standard. Although the primary emphasis in this book is on MPEG video compression, video data streams without audio have very limited value. In addition, without the synchronization and control provided by the MPEG systems layer, the compressed bitstreams are unusable. The basic structure of an MPEG-1 bitstream and the start code syntax are treated here. In addition, a high-level description is given of key aspects of the MPEG video compression algorithms such as the discrete cosine transform (DCT). MPEG-2 and MPEG-4 are briefly discussed.

Chapter 3 discusses the definition and mathematical structure of the 2-D DCT and describes the requirements for an MPEG-compliant DCT. Chapter 4 provides a review of selected aspects of visual perception and applications of this science to MPEG. Chapter 5 provides a review of the basic concepts and principles involved in MPEG compression of video sequences, ending with high-level block diagrams for an MPEG-1 encoder and decoder. Chapter 6 provides a tutorial on the MPEG pseudocode syntax and on the flowchart conventions used in this book. Chapter 7 provides an overview of the MPEG-1 system syntax.

Chapter 8 contains an in-depth discussion of MPEG-1 video bitstream syntax. Both commented pseudocode and flowcharts are presented. Chapter 9 provides a brief overview of the “main profile” of MPEG-2. Chapter 10 describes the syntax of MPEG-2 main profile. Chapter 11 provides an overview of motion compensation techniques, with an emphasis on the block matching techniques relevant to MPEG.

Chapter 12 covers pel reconstruction and includes an in-depth description of IDCT mismatch and the IDCT requirements for MPEG. Chapter 13 describes some of the techniques an encoder may use for estimating motion

displacements. Chapter 14 reviews techniques for varying the DCT quantization in an MPEG encoder. Chapter 15 describes techniques for controlling bitrate.

Chapter 16, written by Cliff Reader, provides a brief analysis of patents that are relevant to MPEG. Chapter 17 provides a sampling of MPEG vendors and products. Chapter 18 gives a history of MPEG, and the final chapter of this book, Chapter 19, briefly describes other standardization efforts in the area of video compression.

1.8 Level of difficulty ○

The reader may have wondered why open circles have been placed after each section heading of this chapter. Since this book is intended for a rather diverse set of readers ranging from nontechnical to highly technical in background, each section is marked with a symbol indicating the approximate level of technical difficulty of the section. The three levels of difficulty are:

- Sections suitable for all readers. For the most part, these sections avoid the use of mathematics and are not dependent on material from more difficult sections.
- Sections of intermediate technical difficulty. These sections build upon the material in the sections marked by ○, and should be understandable to most readers with mathematical or engineering training. They should be of particular interest to MPEG implementors.
- Sections that are either quite challenging or are primarily aimed at those readers who need to delve deeply into the details of MPEG theory or implementation.

1.9 An international collaboration ○

Several hundred technical experts typically attend the MPEG international committee meetings. Many more are involved at their national level. For example, the meetings of the American National Standards Institute (ANSI) committee, X3L3.1, often have more than 100 attendees. At the international level, simultaneous meetings of subgroups concerned with MPEG-1 video, MPEG-1 audio, MPEG-1 systems, MPEG-2 video, MPEG-2 audio, MPEG-2 systems, MPEG-4, testing, etc. are often held, and many companies therefore send more than one representative. With such broad participation, it is virtually impossible to acknowledge each individual contribution.

Many individuals contributed in very significant ways, and the MPEG standard would not have happened without their hard work and dedication and the support of their employers.

The authors of this book have had somewhat different exposures to MPEG during its development. CF and DL have been deeply involved in MPEG. DL chaired the MPEG video subgroup, and CF is a regular participant in MPEG and has been heavily involved in the MPEG software implementation that is Part 5 of the Standard. The other two authors, JM and WP, were occupied with parallel standardization efforts in JPEG and JBIG and did not contribute to the MPEG activities. In some ways these latter two authors felt a bit presumptuous in tackling this book. They started out, in fact, only as editors of this series, but concerted efforts to find alternate authors from the enormous pool of talent contributing to MPEG failed. With the exception of the MPEG video chair, and later, CF, all were too busy either completing the standard or implementing for their companies to be able to meet a tight publication schedule. In some ways, however, having two authors not intimately involved in the generation of the standard has been an advantage, as many things that might be obvious to someone who worked with the standard as it evolved were not so obvious to them.

1.10 An evolving standard ○

Complex standards such as MPEG are not easily arrived at, and require a lot of debate and discussion before consensus is reached. They evolve over time, partly because of a continuing process of refinement and the gradual elimination of competing concepts, and partly because of the influence of new applications that hope to use the standard. Early implementors therefore need to be fully aware that until the standard is fully finished, approved, and published, it may change. MPEG-1 has reached that status, but some parts of MPEG-2 could still be changed. Customers may prefer an early implementation of the Standard that will need upgrading later to maintain compliance rather than a proprietary algorithm, but should be made aware that changes are possible.

One source of information about various MPEG activities that should be current and up-to-date is the Internet World Wide Web MPEG page. The MPEG address is <http://www.mpeg.org> and this is a good place to get background information about the MPEG standards and committee activities. A WWW search should find many sources of MPEG movies, software, and hardware.

Because this standard, in common with most, is still going through a continuing process of refinement, implementors should obtain the latest version and not rely on the contents of either this book or early drafts of the

standard. In the United States, an up-to-date version can be obtained from the American National Standards Institute (ANSI).

Overview of MPEG

ISO/IEC 11172, the document describing MPEG-1, currently contains five parts. Part 1 covers the MPEG system layer, Part 2 describes MPEG video, and Part 3 specifies MPEG audio. Part 4 is on compliance testing, and Part 5 is a software reference model for MPEG-1.

Although the main topic of this book is MPEG video, MPEG audio and MPEG systems are also crucial parts of a complete MPEG implementation. A video bitstream is fully decodable, but, by itself, is an incomplete specification. The MPEG system layer contains control information that enables parsing and precise control of playback of the bitstream. Furthermore, a video stream without audio has only limited use.

This chapter presents an overview of all three parts: systems, audio, and video. As an aid for people who have access to the official MPEG-1 video standard, we sometimes reference a particular clause in that document (clause is standardized for section). These references have the format “IS[w:x.y.z]”, where IS refers to ISO/IEC 11172, w is the part number (systems, video, audio, etc.) and x.y.z is the clause or sub-clause number. If w is 2, it may be omitted. Thus, IS[2:2.3] and IS[2.3] both refer to subclause 2.3 of the video part.

2.1 MPEG system layer ○

The MPEG system layer has the basic task of combining one or more audio and video compressed bitstreams into a single bitstream. It defines the data stream syntax that provides for timing control and the interleaving and synchronization of audio and video bitstreams.

From the systems perspective, an MPEG bitstream is made up of a system layer and compression layers. The system layer provides an envelope for the compression layers. The compression layers contain the data fed

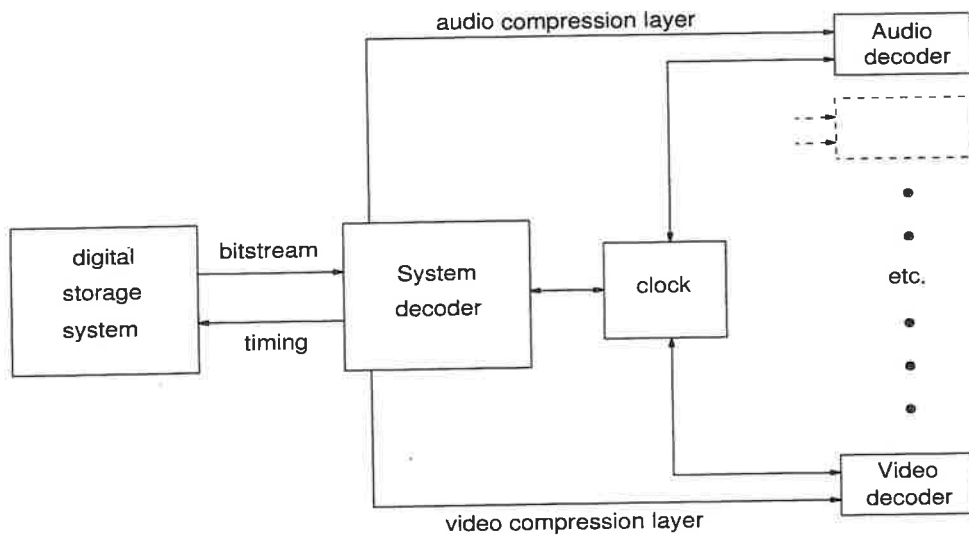


Figure 2.1: MPEG system structure.

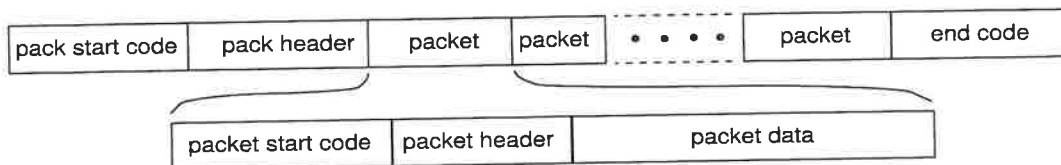


Figure 2.2: System layer pack and packet structure.

to the audio and video decoders, whereas the system layer provides the controls for demultiplexing the interleaved compression layers. A typical MPEG system block diagram is shown in Figure 2.1.

The MPEG bitstream consists of a sequence of packs that are in turn subdivided into packets, as sketched in Figure 2.2. Each pack consists of a unique 32-bit byte-aligned pack start code and header, followed by one or more packets of data. Each packet consists of a packet start code (another unique 32-bit byte-aligned code) and header, followed by packet data (compressed audio or video data). The system decoder parses this bitstream (the pack and packet start codes can be detected without decoding the data) and feeds the separated video and audio data to the appropriate decoders along with timing information.

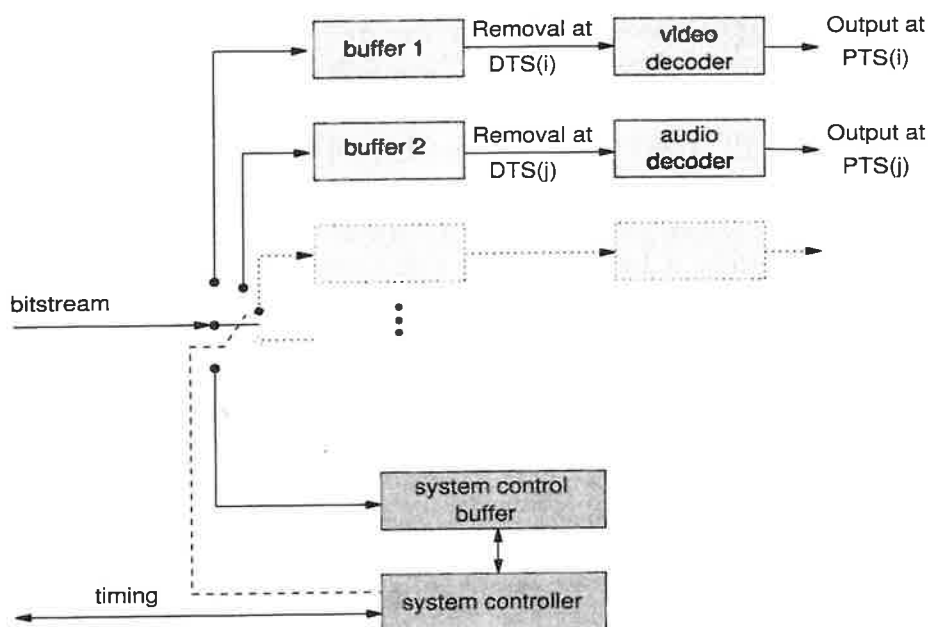


Figure 2.3: A system target decoder with two elementary bitstreams, one video and one audio. The decoders may need to internally reorder the decoded data.

2.1.1 System target decoder ○

The MPEG system uses an idealized decoder called the system target decoder (STD). This idealized decoder interprets the pack and packet headers, delivering the elementary bitstreams to the appropriate audio or video decoder. As shown in Figure 2.1, a number audio and video decoders may be present.

A major role of this idealized decoder is to prevent buffer overflow and underflow. The buffering requirements are described in terms of the decoder rather than the encoder; for this reason, buffer overflow occurs when the decoder does not remove data quickly enough, i.e., when the compression is too good. Conversely, buffer underflow occurs when the decoder removes data too quickly, i.e., when the encoding process produces too much data per picture.

In the STD the bits for an access unit (a picture or an audio access unit) are removed from the buffer instantaneously at a time dictated by a decoding time stamp (DTS) in the bitstream. The bitstream also contains another type of time stamp, the presentation time stamp (PTS). Buffer overflow and underflow is controlled by the DTS; synchronization between audio and

video decoding is controlled by the PTS. At some points in a video sequence, the DTS and PTS are identical and only the PTS is used.

Data are fed to the respective buffers from the system decoder in Figure 2.1 at irregular intervals as needed to demultiplex the interleaved bitstreams. Conceptually, each of the interleaved bitstreams is allocated a separate buffer in the STD, with a size determined by the encoder and communicated to the decoder in the system layer. The situation is sketched in Figure 2.3 for a simple case where one audio and one video bitstream are interleaved.

On the basis of header information in the bitstream itself, the system control knows when to switch the bitstream to buffer 1 or buffer 2. Bits not intended for those buffers are sent instead to the system control buffer. The bits are removed in blocks called *access units*. The DTS time stamps (or PTS, if DTS is not needed) in the bitstream determine the times of transfer and presentation. In this idealized model the transfer, decoding and presentation occur instantaneously; in real systems some additional delays would occur.

2.1.2 System layer syntax ○

An MPEG packet of data has a length defined in the packet header. Packet lengths are typically structured to match to the requirements of the digital storage or transmission medium, and therefore do not necessarily conform to the audio and video access units created by the encoders. However, a packet may only contain one type of compressed data.

2.2 MPEG audio ●

The MPEG audio coding standard defines three layers of increasing complexity and subjective quality.¹ It supports sampling rates of 32, 44.1, and 48 kHz. At 16 bits/sample the uncompressed audio would require about 1.5 Mbit/s. After compression, the bit rates for monophonic channels are between 32 and 192 Kbit/s; the bit rates for stereophonic channels are between 128 and 384 Kbit/s.

The MPEG audio coding techniques take good advantage of the psycho-acoustic properties of human hearing. Just as there is a threshold for observing visible patterns, there is also a frequency-dependent threshold for perception of audio stimuli. There is also an effect known as simultaneous masking in which the presence of one audio signal can mask the perception

¹The MPEG Standard uses the term layers in two different senses — as a descriptive term for the division of the bitstream into MPEG system, audio and video bitstreams, and to distinguish between the three audio algorithms.

of a smaller signal. A temporal masking effect also occurs in which audio signals immediately before and after a masking signal are less perceptible. All of these effects are used to good advantage in the MPEG audio coding models.

The MPEG audio system first segments the audio into windows 384 samples wide. Layers I and II use a filter bank to decompose each window into 32 subbands, each with a width of approximately 750 Hz (for a sampling rate of 48 kHz). As is typical for subband coding, each subband is decimated such that the sampling rate per subband is 1.5 kHz and there are 12 samples per window. A fast Fourier transform (FFT) of the audio input is used to compute a global masking threshold for each subband, and from this a uniform quantizer is chosen that provides the least audible distortion at the required bitrate. Layers I and II are quite similar, but layer II achieves higher performance by using a higher-resolution FFT, finer quantization, and a more efficient way of sending the scale factors for the subbands.

One of the problems introduced by quantization is pre-echoes. Pre-echoes can occur when a sharp percussive sound is preceded by silence. When the signal is reconstructed, the errors due to quantization tend to be distributed over the block of samples, thereby causing an audible distortion before the actual signal. In an 8 ms window, pre-echoes are not suppressed completely by temporal masking.

Pre-echo control is an important part of layer III of MPEG audio coding. Layer III adds a modified discrete cosine transform (DCT) decomposition of the subbands to get much finer frequency subdivision. Layer III also adds nonuniform quantization (larger signals can mask larger quantization errors), entropy coding, and dynamic window switching. Dynamic window switching provides better time resolution, and this allows better control of pre-echoes.

A recent review article by Pan [Pan94] gives more details on MPEG audio coding. The book by Bhaskaran and Konstantinides [BK95] has reviews of many recent standards, including MPEG audio.

2.3 MPEG video ○

MPEG video is specifically designed for compression of video sequences. A video sequence is simply a series of pictures taken at closely spaced intervals in time. Except for the special case of a scene change, these pictures tend to be quite similar from one to the next. Intuitively, a compression system ought to be able to take advantage of this similarity.

It should come as no surprise to the reader, therefore, that the compression techniques (*compression models*) used by MPEG take advantage of this similarity or predictability from one picture to the next in a sequence. Com-

pression techniques that use information from other pictures in the sequence are usually called interframe techniques.

When a scene change occurs (and sometimes for other reasons), interframe compression does not work and the compression model should be changed. In this case the compression model should be structured to take advantage of the similarity of a given region of a picture to immediately adjacent areas in the same picture. Compression techniques that only use information from a single picture are usually called intraframe techniques.

These two compression techniques, interframe and intraframe, are at the heart of the MPEG video compression algorithm. However, because there might be confusion with the term *frame* as used in broadcast video systems, MPEG-1 uses the abbreviated terms, *inter* and *intra*. In addition, MPEG very often uses the term *non-intra* in place of *inter*. These definitions occur in IS[2.1.78, 2.1.79, and 2.1.95]. We will follow the MPEG-1 convention for the rest of this book.

2.3.1 MPEG video layers ○

The outermost layer of an MPEG video bitstream is the video sequence layer. Except for certain critical timing information in the MPEG systems layer, an MPEG video sequence bitstream is completely self-contained. It is independent of other video (and audio) bitstreams.

Each video sequence is divided into one or more groups of pictures, and each group of pictures is composed of one or more pictures of three different types, I-, P-, and B-, as illustrated in Figure 2.4.² I-pictures (intra-coded pictures) are coded independently, entirely without reference to other pictures. P- and B-pictures are compressed by coding the differences between the picture and reference I- or P-pictures, thereby exploiting the similarities from one picture to the next.

P-pictures (predictive-coded pictures) obtain predictions from temporally preceding I- or P-pictures in the sequence, whereas B-pictures (bidirectionally predictive-coded pictures) obtain predictions from the nearest preceding and/or upcoming I- or P-pictures in the sequence. Different regions of B-pictures may use different predictions, and may predict from preceding pictures, upcoming pictures, both, or neither. Similarly, P-pictures may also predict from preceding pictures or use no prediction. If no prediction is used, that region of the picture is coded by intra techniques.

In a closed group of pictures P- and B-pictures are predicted only from other pictures in the group of pictures; in an open group of pictures the prediction may be from pictures outside of the group of pictures.

²A fourth type, the D-picture, is defined but is rarely used. The D-picture is a low-resolution representation that may not be used in combination with I-, P-, or B-pictures.

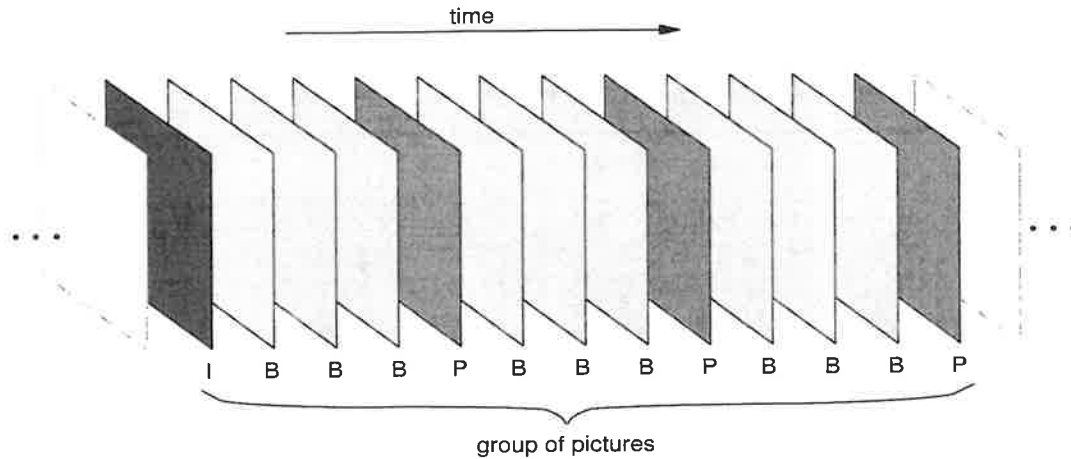


Figure 2.4: A typical group of pictures in display order.

2.3.2 Display and coding order ○

Since MPEG sometimes uses information from future pictures in the sequence, the coding order, the order in which compressed pictures are found in the bitstream, is not the same as the display order, the order in which pictures are presented to a viewer. The coding order is the order in which the pictures should be decoded by the decoder. The group of pictures illustrated in Figure 2.4 is shown in coding order in Figure 2.5.

2.3.3 Macroblock ○

The basic building block of an MPEG picture is the macroblock, sketched in Figure 2.6. The macroblock consists of a 16×16 sample array of luminance (grayscale) samples together with one 8×8 block of samples for each of two chrominance (color) components. The 16×16 sample array of luminance samples is actually composed of four 8×8 blocks of samples, and these 8×8 blocks are the units of data that are fed to the compression models. The reason for this will become clearer in the discussion of the coding models.

2.3.4 Slice ○

The MPEG picture is not simply an array of macroblocks, however. Rather, as sketched in Figure 2.7, it is composed of slices, where each slice is a contiguous sequence of macroblocks in raster scan order, starting at a specific address or position in the picture specified in the slice header. Each small

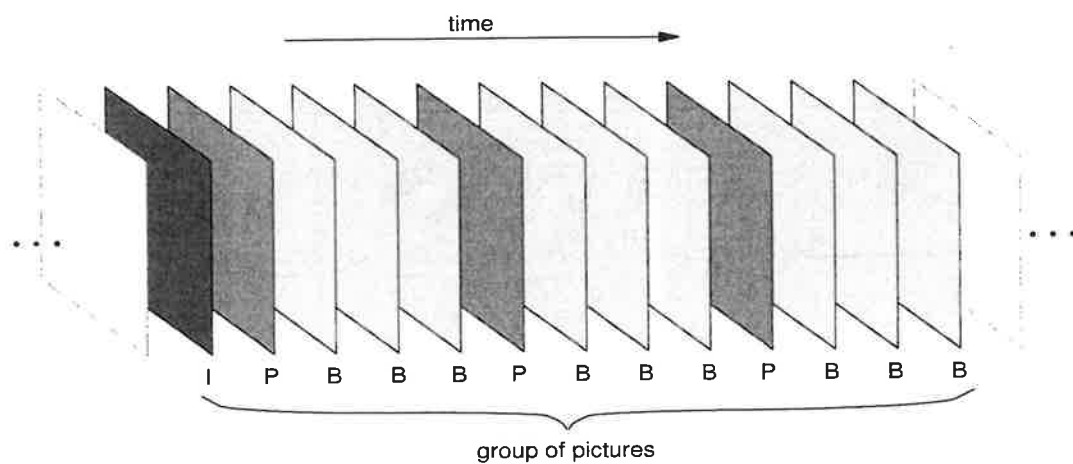


Figure 2.5: A typical group of pictures in coding order.

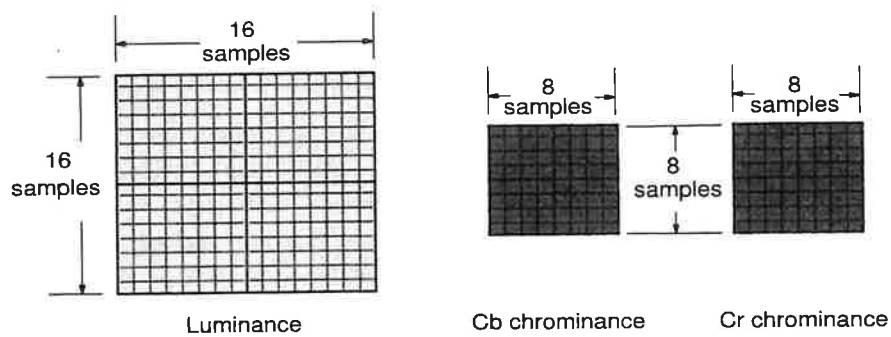


Figure 2.6: The MPEG macroblock.

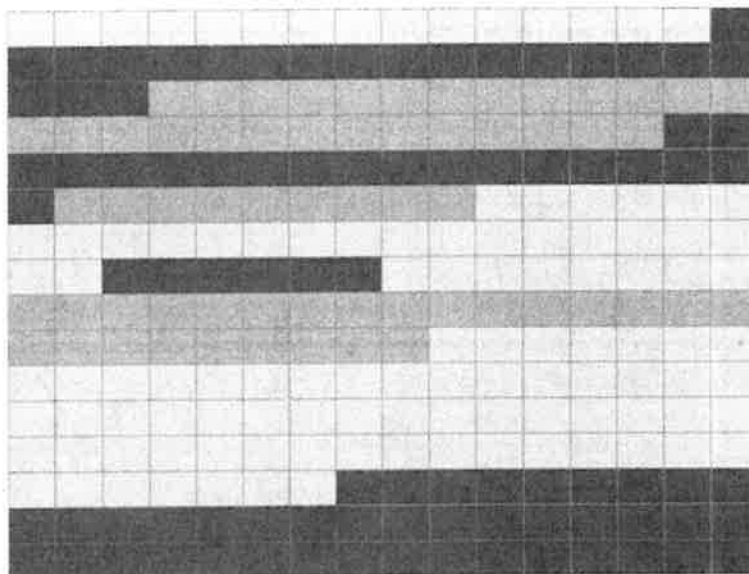


Figure 2.7: An illustration of a possible slice structure in an MPEG-1 picture.

block in the figure represents a macroblock, and contiguous macroblocks in a given slice have the same shade of gray. Slices can continue from one macroblock row to the next in MPEG-1. This slice structure, among other things, allows for great flexibility in signaling changes in some of the coding parameters. This is needed both to optimize quality for a given bitrate and to control that bitrate.

2.3.5 The discrete cosine transform in MPEG ○

At the heart of both intra and inter coding in MPEG is the discrete cosine transform (DCT). The DCT has certain properties that simplify coding models and make the coding efficient in terms of perceptual quality measures.

Basically, the DCT is a method of decomposing a block of data into a weighted sum of spatial frequencies. Figure 2.8 illustrates the spatial frequency patterns that are used for an 8×8 DCT (a DCT for an 8×8 block of pels). Each of these spatial frequency patterns has a corresponding *coefficient*, the amplitude needed to represent the contribution of that spatial frequency pattern in the block of data being analyzed. In other words, each spatial frequency pattern is multiplied by its coefficient and the resulting 64

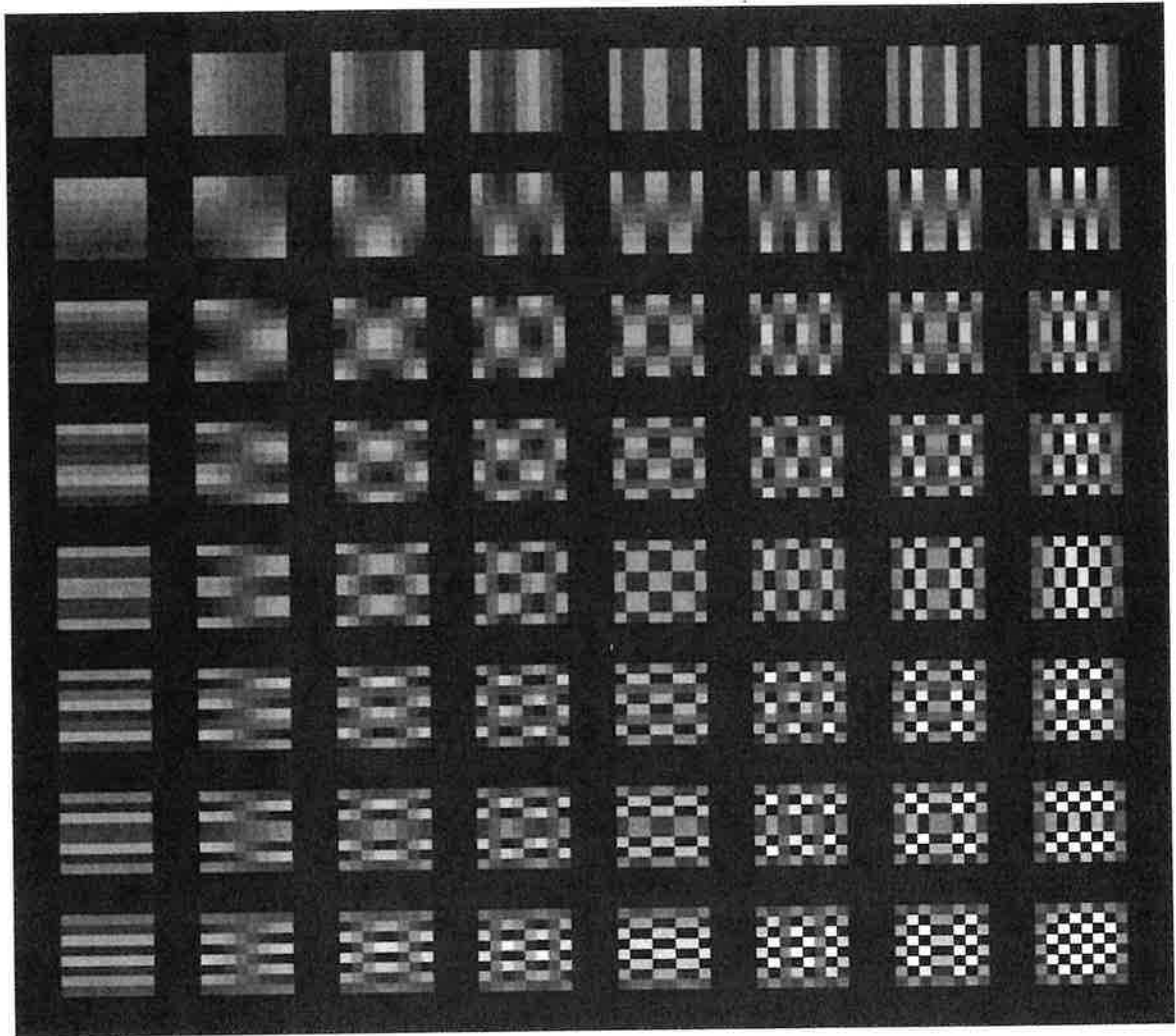


Figure 2.8: The 64 2-D cosine functions of the 8×8 DCT. Each block is an 8×8 array of samples. Zero amplitudes are neutral gray, negative amplitudes have darker intensities and positive amplitudes have lighter intensities.

8×8 amplitude arrays are summed, each pel separately, to reconstruct the 8×8 block.

If only the low-frequency DCT coefficients are nonzero, the data in the block vary slowly with position. If high frequencies are present, the block intensity changes rapidly from pel to pel. This qualitative behavior can be seen in Figure 2.8, where the 64 two-dimensional (2-D) patterns that make up the 8×8 DCT are seen to range from absolutely flat — the term that gives the average or “DC value” of the 8×8 block of pels — to a checkerboard with very large intensity changes from one pel to the next.³ Note that at each axis of the plot, the variation is one-dimensional. As will be shown in Chapter 3, the complex 2-D behavior of Figure 2.8 is actually the product of the one-dimensional (1-D) oscillations seen along each axis. Qualitatively, this can be seen by closely inspecting Figure 2.8.

2.3.6 Quantization ○

When the DCT is computed for a block of pels, it is desirable to represent the coefficients for high spatial frequencies with less precision. This is done by a process called *quantization*. A DCT coefficient is quantized by dividing it by a nonzero positive integer called a quantization value and rounding the quotient — the quantized DCT coefficient — to the nearest integer. The bigger the quantization value is, the lower the precision is of the quantized DCT coefficient. Lower-precision coefficients can be transmitted to a decoder with fewer bits. The use of large quantization values for high spatial frequencies allows the encoder to selectively discard high spatial frequency activity that the human eye cannot readily perceive.

The DCT and visually-weighted quantization of the DCT are key parts of the MPEG coding system. In Chapter 3 we will discuss the DCT in some detail; in Chapter 4 we will explore properties of the human visual system that relate to quantization and other aspects of MPEG coding.

As noted above, a macroblock is composed of four 8×8 blocks of luminance (monochrome) samples and two 8×8 blocks of chrominance samples. The chrominance samples represent color in terms of the presence or absence of red and blue for a given luminance intensity. These 8×8 blocks of data are the unit processed by the DCT. A lower resolution is used for the chrominance blocks because the human eye resolves high spatial frequencies in luminance better than in chrominance. This will be covered in detail in Chapter 4.

The DCT turns out to have several advantages from the point of view of data compression. First, it turns out that for intra coding, the DCT

³The term *DC* stands for direct current, and comes from the use of the DCT to analyze electrical circuits. Similarly, currents varying with time are called *AC* (alternating current) terms; in our case, AC applies to spatially varying terms.

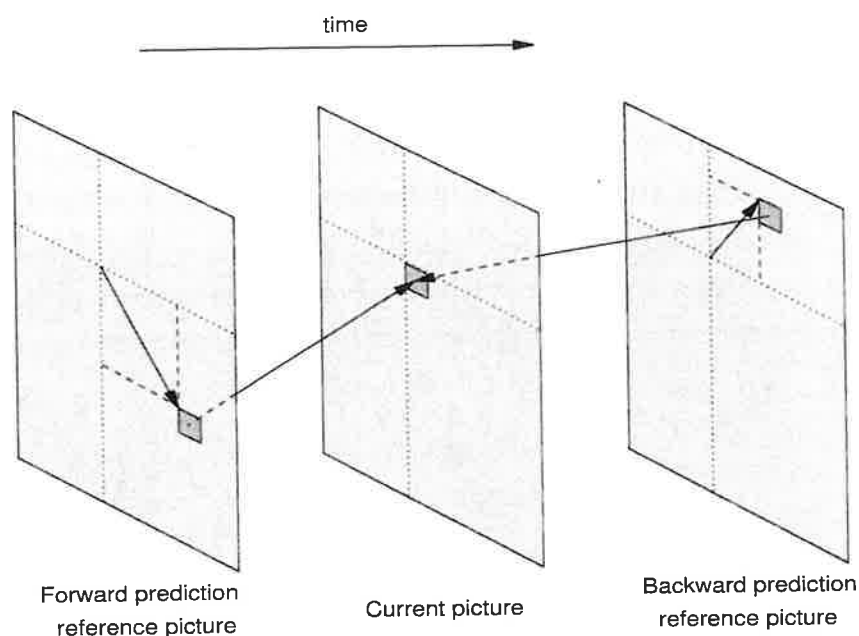


Figure 2.9: Motion compensated prediction and reconstruction.

coefficients are almost completely decorrelated — that is, they are independent of one another, and therefore can be coded independently. This makes it possible to design a relatively simple algorithm (called a coding model) for coding them. Decorrelation is of great theoretical and practical interest in terms of construction of the coding model. However, the coding performance is actually influenced much more profoundly by the visually-weighted quantization.

In nonintra coding — coding the difference between the current picture and a picture already transmitted — the DCT does not greatly improve the decorrelation, since the difference signal obtained by subtracting the prediction from a similar (i.e., correlated) picture is already fairly well decorrelated. However, quantization is still a powerful compression for controlling bitrate, even if decorrelation is not improved very much by the DCT. In practice, the DCT works quite well for both inter and intra coding.

2.3.7 Motion compensation ○

If there is motion in the sequence, a better prediction is often obtained by coding differences relative to areas that are shifted with respect to the area being coded, a process known as motion compensation. The process of determining the motion vectors in the encoder is called motion estimation.

The motion vectors describing the direction and amount of motion of



2.3.8 MPEG coding models○

The coefficients are arranged according to a 1-D sequence known as the zigzag scanning order. This zigzag ordering is shown in Figure 2.10, where we see that this scan approximately orders the coefficients in ascending spatial frequency. Since visually-weighted quantization strongly deemphasizes higher spatial frequencies, only a few lower-frequency coefficients are nonzero in a typical transformation. As a rough rule-of-thumb, the more zero coefficients, the better the compression. Consequently, it is very important

for the coding model to use *symbols* (particular combinations of DCT coefficient information) that permit efficient coding of DCTs with mostly zero coefficients. One of these symbols is an end-of-block (EOB), that tells the decoder that all remaining DCT coefficients for the block are zero. More details about the coding model will be covered in Chapter 5.

The DCT is used to code both nonintra and intra picture information, but the properties are actually quite different in the two coding environments. Different quantization tables are used for nonintra compression, and the rounding is done differently. The data being transformed are also quite different, and therefore, different sets of symbols and slightly different code tables are used.

Many other aspects of MPEG, such as the methods for simply skipping over unchanged macroblocks and for coding of motion vectors, will be covered in Chapter 8 on MPEG video syntax. While it is inappropriate to get into such detail in this overview, these details are important in determining the overall coding performance that can be achieved by MPEG.

2.3.9 Encoder decision strategies ○

It is important to recognize that MPEG is potentially a highly asymmetric system. Decoders, since they merely follow the directions encoded in the bitstream, are relatively simple. Encoders, however, are much more complex than decoders and must have much more intelligence. Among other things, encoders must identify areas in motion, determine optimal motion vectors, control bitrate, control data buffering such that underflow and overflow do not occur, determine where to change quantization, determine when a given block can simply be repeated, determine when to code by intra and inter techniques, and vary all of these parameters and decisions dynamically so as to maximize quality for a given rate. The decision strategy that must be implemented in the encoder is probably the most complex and least published aspect of MPEG.

2.4 MPEG-2 ○

MPEG-2 is the follow-on to MPEG-1, and is intended primarily for higher bit rates, larger picture sizes, and interlaced frames. MPEG-2 builds upon the MPEG-1 capabilities. In fact, MPEG-2 decoders are required to be able to decode at least some MPEG-1 data streams.

Five different profiles have been defined for MPEG-2 video, providing particular sets of capabilities ranging from function close to MPEG-1 to the very advanced video compression techniques needed for HDTV. One of the profiles, the main profile (MP) of MPEG-2, is, however, a relatively

straight-forward extension of MPEG-1. The main profile of MPEG-2 video is covered in this book. Briefly, MPEG-2 main profile introduces the concept of video fields and interlaced frames. Variable quantization is extended, DC precision can be customized, and another coding table is defined. Higher resolutions, bit rates, and frame rates are expected. Parameter ranges such as picture width, height, bit rate, and buffer sizes are extended. Copyright information can be recorded. Display characteristics can be transmitted including an offset for viewing a subset of the picture in a window.

MPEG-2 systems (Part 1) includes all of the features of MPEG-1 with enhancements that allow it to operate in more error-prone environments. In addition, more video and audio streams can be handled and they are no longer required to have a common time base. MPEG-2 systems still uses the pack and packet concepts, but combines them in both backwards compatible and noncompatible ways. Some of the noncompatible ways support asynchronous transmission mode (ATM) applications of MPEG-2.

MPEG-2 audio (Part 3) is backwards compatible with MPEG-1. Even though three more channels are allowed plus a sub-woofer to provide an enhanced surround sound environment, an MPEG-1 audio decoder can extract and decode two audio channels. Also, an MPEG-2 audio decoder can decode the two stereo channels of MPEG-1 audio. The requirement to be backwards compatible necessitated some coding efficiency compromises. A nonbackwards compatible (NBC) audio (Part 7) is also being developed.

MPEG-1 video was aimed at the digital storage media (DSM) application and thus fixed the video format (see Figure 1.3), assumed low errors rates, and ignored interlaced TV. MPEG-2 was intended to be a generic standard and therefore attempted to be much broader. MPEG-2 video (Part 2) maintains all of the MPEG-1 video syntax, but uses extensions to add additional flexibility and functions. “Scalable” extensions are added to provide video data streams with multiple resolutions for potential normal TV and HDTV coexistence. Other scalable extensions allow the data stream to be partitioned into two pieces. Others also offer some temporal flexibility so that not all frames have to be reconstructed.

2.5 MPEG-4 ○

Before MPEG-2 was finished a new project, MPEG-4, was started.⁴ It was originally targeted primarily at very low bit rates, but over time its focus changed beyond just high compression. High priority was placed on content-based interactivity and “universal access”, including error-prone wireless networks [ISO95]. Content-based manipulation, scalability, and editing are

⁴MPEG-3 was started and targeted at HDTV, but abandoned when MPEG-2 was demonstrated to meet that need.

also expected [Rea96]⁵ MPEG-4 will provide a syntactic description language (MSDL). This MSDL is expected to describe how to parse and process the elementary data streams (which could be MPEG-1, MPEG-2, or even synthetic two-dimensional or three-dimensional objects). Missing pieces will be downloaded from an open set of tools and algorithms.

⁵A special issue on MPEG-4 is scheduled for the December 1996 issue of the IEEE Transactions on Circuits and Systems for Video Technology.