

Iain E. G. Richardson

H.264 and MPEG-4

VIDEO COMPRESSION



 **WILEY**



H.264 and MPEG-4 Video Compression

H.264 and MPEG-4 Video Compression

Video Coding for Next-generation Multimedia

Iain E. G. Richardson

The Robert Gordon University, Aberdeen, UK



Copyright © 2003 John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England
Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wileyeurope.com or www.wiley.com

All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1T 4LP, UK, without the permission in writing of the Publisher.

Requests to the Publisher should be addressed to the Permissions Department, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, or emailed to permreq@wiley.co.uk, or faxed to (+44) 1243 770620.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0-470-84837-5

Typeset in 10/12pt Times roman by TechBooks, New Delhi, India

Printed and bound in Great Britain by Antony Rowe, Chippenham, Wiltshire

This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

Contents

About the Author	xiii
Foreword	xv
Preface	xix
Glossary	xxi
1 Introduction	1
1.1 The Scene	1
1.2 Video Compression	3
1.3 MPEG-4 and H.264	5
1.4 This Book	6
1.5 References	7
2 Video Formats and Quality	9
2.1 Introduction	9
2.2 Natural Video Scenes	9
2.3 Capture	10
2.3.1 Spatial Sampling	11
2.3.2 Temporal Sampling	11
2.3.3 Frames and Fields	13
2.4 Colour Spaces	13
2.4.1 RGB	14
2.4.2 YCbCr	15
2.4.3 YCbCr Sampling Formats	17
2.5 Video Formats	19
2.6 Quality	20
2.6.1 Subjective Quality Measurement	21
2.6.2 Objective Quality Measurement	22
2.7 Conclusions	24
2.8 References	24

3 Video Coding Concepts	27
3.1 Introduction	27
3.2 Video CODEC	28
3.3 Temporal Model	30
3.3.1 Prediction from the Previous Video Frame	30
3.3.2 Changes due to Motion	30
3.3.3 Block-based Motion Estimation and Compensation	32
3.3.4 Motion Compensated Prediction of a Macroblock	33
3.3.5 Motion Compensation Block Size	34
3.3.6 Sub-pixel Motion Compensation	37
3.3.7 Region-based Motion Compensation	41
3.4 Image model	42
3.4.1 Predictive Image Coding	44
3.4.2 Transform Coding	45
3.4.3 Quantisation	51
3.4.4 Reordering and Zero Encoding	56
3.5 Entropy Coder	61
3.5.1 Predictive Coding	61
3.5.2 Variable-length Coding	62
3.5.3 Arithmetic Coding	69
3.6 The Hybrid DPCM/DCT Video CODEC Model	72
3.7 Conclusions	82
3.8 References	83
4 The MPEG-4 and H.264 Standards	85
4.1 Introduction	85
4.2 Developing the Standards	85
4.2.1 ISO MPEG	86
4.2.2 ITU-T VCEG	87
4.2.3 JVT	87
4.2.4 Development History	88
4.2.5 Deciding the Content of the Standards	88
4.3 Using the Standards	89
4.3.1 What the Standards Cover	90
4.3.2 Decoding the Standards	90
4.3.3 Conforming to the Standards	91
4.4 Overview of MPEG-4 Visual/Part 2	92
4.5 Overview of H.264 / MPEG-4 Part 10	93
4.6 Comparison of MPEG-4 Visual and H.264	94
4.7 Related Standards	95
4.7.1 JPEG and JPEG2000	95
4.7.2 MPEG-1 and MPEG-2	95
4.7.3 H.261 and H.263	96
4.7.4 Other Parts of MPEG-4	97
4.8 Conclusions	97
4.9 References	98

5 MPEG-4 Visual	99
5.1 Introduction	99
5.2 Overview of MPEG-4 Visual (Natural Video Coding)	100
5.2.1 Features	100
5.2.2 Tools, Objects, Profiles and Levels	100
5.2.3 Video Objects	103
5.3 Coding Rectangular Frames	104
5.3.1 Input and Output Video Format	106
5.3.2 The Simple Profile	106
5.3.3 The Advanced Simple Profile	115
5.3.4 The Advanced Real Time Simple Profile	121
5.4 Coding Arbitrary-shaped Regions	122
5.4.1 The Core Profile	124
5.4.2 The Main Profile	133
5.4.3 The Advanced Coding Efficiency Profile	138
5.4.4 The N-bit Profile	141
5.5 Scalable Video Coding	142
5.5.1 Spatial Scalability	142
5.5.2 Temporal Scalability	144
5.5.3 Fine Granular Scalability	145
5.5.4 The Simple Scalable Profile	148
5.5.5 The Core Scalable Profile	148
5.5.6 The Fine Granular Scalability Profile	149
5.6 Texture Coding	149
5.6.1 The Scalable Texture Profile	152
5.6.2 The Advanced Scalable Texture Profile	152
5.7 Coding Studio-quality Video	153
5.7.1 The Simple Studio Profile	153
5.7.2 The Core Studio Profile	155
5.8 Coding Synthetic Visual Scenes	155
5.8.1 Animated 2D and 3D Mesh Coding	155
5.8.2 Face and Body Animation	156
5.9 Conclusions	156
5.10 References	156
6 H.264/MPEG-4 Part 10	159
6.1 Introduction	159
6.1.1 Terminology	159
6.2 The H.264 CODEC	160
6.3 H.264 structure	162
6.3.1 Profiles and Levels	162
6.3.2 Video Format	162
6.3.3 Coded Data Format	163
6.3.4 Reference Pictures	163
6.3.5 Slices	164
6.3.6 Macroblocks	164

6.4 The Baseline Profile	165
6.4.1 Overview	165
6.4.2 Reference Picture Management	166
6.4.3 Slices	167
6.4.4 Macroblock Prediction	169
6.4.5 Inter Prediction	170
6.4.6 Intra Prediction	177
6.4.7 Deblocking Filter	184
6.4.8 Transform and Quantisation	187
6.4.9 4×4 Luma DC Coefficient Transform and Quantisation (16 $\times$ 16 Intra-mode Only)	194
6.4.10 2×2 Chroma DC Coefficient Transform and Quantisation	195
6.4.11 The Complete Transform, Quantisation, Rescaling and Inverse Transform Process	196
6.4.12 Reordering	198
6.4.13 Entropy Coding	198
6.5 The Main Profile	207
6.5.1 B Slices	207
6.5.2 Weighted Prediction	211
6.5.3 Interlaced Video	212
6.5.4 Context-based Adaptive Binary Arithmetic Coding (CABAC)	212
6.6 The Extended Profile	216
6.6.1 SP and SI slices	216
6.6.2 Data Partitioned Slices	220
6.7 Transport of H.264	220
6.8 Conclusions	222
6.9 References	222
7 Design and Performance	225
7.1 Introduction	225
7.2 Functional Design	225
7.2.1 Segmentation	226
7.2.2 Motion Estimation	226
7.2.3 DCT/IDCT	234
7.2.4 Wavelet Transform	238
7.2.5 Quantise/Rescale	238
7.2.6 Entropy Coding	238
7.3 Input and Output	241
7.3.1 Interfacing	241
7.3.2 Pre-processing	242
7.3.3 Post-processing	243
7.4 Performance	246
7.4.1 Criteria	246
7.4.2 Subjective Performance	247
7.4.3 Rate–distortion Performance	251

7.4.4 Computational Performance	254
7.4.5 Performance Optimisation	255
7.5 Rate control	256
7.6 Transport and Storage	262
7.6.1 Transport Mechanisms	262
7.6.2 File Formats	263
7.6.3 Coding and Transport Issues	264
7.7 Conclusions	265
7.8 References	265
8 Applications and Directions	269
8.1 Introduction	269
8.2 Applications	269
8.3 Platforms	270
8.4 Choosing a CODEC	270
8.5 Commercial issues	272
8.5.1 Open Standards?	273
8.5.2 Licensing MPEG-4 Visual and H.264	274
8.5.3 Capturing the Market	274
8.6 Future Directions	275
8.7 Conclusions	276
8.8 References	276
Bibliography	277
Index	279

4

The MPEG-4 and H.264 Standards

4.1 INTRODUCTION

An understanding of the process of creating the standards can be helpful when interpreting or implementing the documents themselves. In this chapter we examine the role of the ISO MPEG and ITU VCEG groups in developing the standards. We discuss the mechanisms by which the features and parameters of the standards are chosen and the driving forces (technical and commercial) behind these mechanisms. We explain how to ‘decode’ the standards and extract useful information from them and give an overview of the two standards covered by this book, MPEG-4 Visual (Part 2) [1] and H.264/MPEG-4 Part 10 [2]. The technical scope and details of the standards are presented in Chapters 5 and 6 and in this chapter we concentrate on the target applications, the ‘shape’ of each standard and the method of specifying the standard. We briefly compare the two standards with related International Standards such as MPEG-2, H.263 and JPEG.

4.2 DEVELOPING THE STANDARDS

Creating, maintaining and updating the ISO/IEC 14496 (‘MPEG-4’) set of standards is the responsibility of the Moving Picture Experts Group (MPEG), a study group who develop standards for the International Standards Organisation (ISO). The emerging H.264 Recommendation (also known as MPEG-4 Part 10, ‘Advanced Video Coding’ and formerly known as H.26L) is a joint effort between MPEG and the Video Coding Experts Group (VCEG), a study group of the International Telecommunications Union (ITU).

MPEG developed the highly successful MPEG-1 and MPEG-2 standards for coding video and audio, now widely used for communication and storage of digital video, and is also responsible for the MPEG-7 standard and the MPEG-21 standardisation effort. VCEG was responsible for the first widely-used videotelephony standard (H.261) and its successor, H.263, and initiated the early development of the H.26L project. The two groups set-up the collaborative Joint Video Team (JVT) to finalise the H.26L proposal and convert it into an international standard (H.264/MPEG-4 Part 10) published by both ISO/IEC and ITU-T.

H.264 and MPEG-4 Video Compression: Video Coding for Next-generation Multimedia.
Iain E. G. Richardson. © 2003 John Wiley & Sons, Ltd. ISBN: 0-470-84837-5

Table 4.1 MPEG sub-groups and responsibilities [5]

Sub-group	Responsibilities
Requirements	Identifying industry needs and requirements of new standards.
Systems	Combining audio, video and related information; carrying the combined data on delivery mechanisms.
Description	Declaring and describing digital media items.
Video	Coding of moving images.
Audio	Coding of audio.
Synthetic Natural Hybrid Coding	Coding of synthetic audio and video for integration with natural audio and video.
Integration	Conformance testing and reference software.
Test	Methods of subjective quality assessment.
Implementation	Experimental frameworks, feasibility studies, implementation guidelines.
Liaison	Relations with other relevant groups and bodies.
Joint Video Team	See below.

4.2.1 ISO MPEG

The Moving Picture Experts Group is a Working Group of the International Organisation for Standardization (ISO) and the International Electrotechnical Commission (IEC). Formally, it is Working Group 11 of Subcommittee 29 of Joint Technical Committee 1 and hence its official title is ISO/IEC JTC1/SC29/WG11.

MPEG's remit is to develop standards for compression, processing and representation of moving pictures and audio. It has been responsible for a series of important standards starting with MPEG-1 (compression of video and audio for CD playback) and following on with the very successful MPEG-2 (storage and broadcasting of 'television-quality' video and audio). MPEG-4 (coding of audio-visual objects) is the latest standard that deals specifically with audio-visual coding. Two other standardisation efforts (MPEG-7 [3] and MPEG-21 [4]) are concerned with multimedia content representation and a generic 'multimedia framework' respectively¹. However, MPEG is arguably best known for its contribution to audio and video compression. In particular, MPEG-2 is ubiquitous at the present time in digital TV broadcasting and DVD-Video and MPEG Layer 3 audio coding ('MP3') has become a highly popular mechanism for storage and sharing of music (not without controversy).

MPEG is split into sub-groups of experts, each addressing a particular issue related to the standardisation efforts of the group (Table 4.1). The 'Experts' in MPEG are drawn from a diverse, world-wide range of companies, research organisations and institutes. Membership of MPEG and participation in its meetings is restricted to delegates of National Standards Bodies. A company or institution wishing to participate in MPEG may apply to join its National Standards Body (for example, in the UK this is done through committee IST/37 of the British Standards Institute). Joining the standards body, attending national meetings and attending the international MPEG meetings can be an expensive and time consuming exercise.

¹ So what happened to the missing numbers? MPEG-3 was originally intended to address coding for high-definition video but this was incorporated into the scope of MPEG-2. By this time, MPEG-4 was already under development and so '3' was missed out. The sequential numbering system was abandoned for MPEG-7 and MPEG-21.

However, the benefits include access to the private documents of MPEG (including access to the standards in draft form before their official publication, providing a potential market lead over competitors) and the opportunity to shape the development of the standards.

MPEG conducts its business at meetings that take place every 2–3 months. Since 1988 there have been 64 meetings (the first was in Ottawa, Canada and the latest in Pattaya, Thailand). A meeting typically lasts for five days, with ad hoc working groups meeting prior to the main meeting. At the meeting, proposals and developments are deliberated by each sub-group, successful contributions are approved and, if appropriate, incorporated into the standardisation process. The business of each meeting is recorded by a set of input and output documents. Input documents include reports by ad hoc working groups, proposals for additions or changes to the developing standards and statements from other bodies and output documents include meeting reports and draft versions of the standards.

4.2.2 ITU-T VCEG

The Video Coding Experts Group is a working group of the International Telecommunication Union Telecommunication Standardisation Sector (ITU-T). ITU-T develops standards (or ‘Recommendations’) for telecommunication and is organised in sub-groups. Sub-group 16 is responsible for multimedia services, systems and terminals, working party 3 of SG16 addresses media coding and the work item that led to the H.264 standard is described as Question 6. The official title of VCEG is therefore ITU-T SG16 Q.6. The name VCEG was adopted relatively recently and previously the working group was known by its ITU designation and/or as the Low Bitrate Coding (LBC) Experts Group.

VCEG has been responsible for a series of standards related to video communication over telecommunication networks and computer networks. The H.261 videoconferencing standard was followed by the more efficient H.263 which in turn was followed by later versions (informally known as H.263+ and H.263++) that extended the capabilities of H.263. The latest standardisation effort, previously known as ‘H.26L’, has led to the development and publication of Recommendation H.264.

Since 2001, this effort has been carried out cooperatively between VCEG and MPEG and the new standard, entitled ‘Advanced Video Coding’ (AVC), is jointly published as ITU-T H.264 and ISO/IEC MPEG-4 Part 10.

Membership of VCEG is open to any interested party (subject to approval by the chairman). VCEG meets at approximately three-monthly intervals and each meeting considers a series of objectives and proposals, some of which are incorporated into the developing draft standard. Ad hoc groups are formed in order to investigate and report back on specific problems and questions and discussion is carried out via email reflector between meetings. In contrast with MPEG, the input and output documents of VCEG (and now JVT, see below) are publicly available. Earlier VCEG documents (from 1996 to 1992) are available at [6] and documents since May 2002 are available by FTP [7].

4.2.3 JVT

The Joint Video Team consists of members of ISO/IEC JTC1/SC29/WG11 (MPEG) and ITU-T SG16 Q.6 (VCEG). JVT came about as a result of an MPEG requirement for advanced video coding tools. The core coding mechanism of MPEG-4 Visual (Part 2) is based on rather

Table 4.2 MPEG-4 and H.264 development history

1993	MPEG-4 project launched. Early results of H. 263 project produced.
1995	MPEG-4 call for proposals including efficient video coding and content-based functionalities. H.263 chosen as core video coding tool
1998	Call for proposals for H.26L.
1999	MPEG-4 Visual standard published. Initial Test Model (TM1) of H.26L defined.
2000	MPEG call for proposals for advanced video coding tools.
2001	Edition 2 of the MPEG-4 Visual standard published. H.26L adopted as basis for proposed MPEG-4 Part 10. JVT formed.
2002	Amendments 1 and 2 (Studio and Streaming Video profiles) to MPEG-4 Visual Edition 2 published. H.264 technical content frozen.
2003	H.264/MPEG-4 Part 10 ('Advanced Video Coding') published.

old technology (the original H.263 standard, published as a standard in 1995). With recent advances in processor capabilities and video coding research it was clear that there was scope for a step-change in video coding performance. After evaluating several competing technologies in June 2001, it became apparent that the H.26L test model CODEC was the best choice to meet MPEG's requirements and it was agreed that members of MPEG and VCEG would form a Joint Video Team to manage the final stages of H.26L development. Under its Terms of Reference [8], JVT meetings are co-located with meetings of VCEG and/or MPEG and the outcomes are reported back to the two parent bodies. JVT's main purpose was to see the H.264 Recommendation / MPEG-4 Part 10 standard through to publication ; now that the standard is complete, the group's focus has switched to extensions to support other colour spaces and increased sample accuracy.

4.2.4 Development History

Table 4.2 lists some of the major milestones in the development of MPEG-4 Visual and H.264. The focus of MPEG-4 was originally to provide a more flexible and efficient update to the earlier MPEG-1 and MPEG-2 standards. When it became clear in 1994 that the newly-developed H.263 standard offered the best compression technology available at the time, MPEG changed its focus and decided to embrace object-based coding and functionality as the distinctive element of the new MPEG-4 standard. Around the time at which MPEG-4 Visual was finalised (1998/99), the ITU-T study group began evaluating proposals for a new video coding initiative entitled H.26L (the L stood for 'Long Term'). The developing H.26L test model was adopted as the basis for the proposed Part 10 of MPEG-4 in 2001.

4.2.5 Deciding the Content of the Standards

Development of a standard for video coding follows a series of steps. A work plan is agreed with a set of functional and performance objectives for the new standard. In order to decide upon the basic technology of the standard, a competitive trial may be arranged in which interested parties are invited to submit their proposed solution for evaluation. In the case of video and image coding, this may involve evaluating the compression and processing performance of a number of software CODECs. In the cases of MPEG-4 Visual and H.264, the chosen core technology for video coding is based on block-based motion compensation followed by transformation and quantisation of the residual. Once the basic algorithmic core of the standard has been decided,

much further work is required to develop the detail of the standard. Proposals for algorithms, methods or functionalities are made by companies or organisations with an interest in the outcome of the standardisation effort and are submitted to the standardisation group for evaluation. At each meeting of the standards group, proposal submissions are discussed and may be adopted or discarded. Through this process, the detail of the standard is gradually refined.

A software model ('reference software') is developed that implements the current agreed functionality, together with a document describing this software ('test model'). The reference software and test model are revised and updated to incorporate each agreed refinement to the standard. Eventually, the standards group decides that the test model document is at a sufficiently mature stage and it is converted into a draft standard. After further refinement and a series of ballots, the document is published as an International Standard.

The content of the final International Standard is reached through an organic development process that involves hundreds of participants (the members of the standardisation group) and many modifications and compromises. The final document is shaped by a number of driving forces. On the technical side, the contributors to the MPEG and VCEG meetings are drawn from the leading research institutes, universities and companies in the video coding field. The standards groups attempt to incorporate the best of current research and development in order to achieve the desired compression and functional performance within practical computational constraints. On the commercial side, the initial work plan of the standardisation group aims to anticipate the technology requirements of future applications. Technical proposals have to be evaluated in terms of cost-effective implementation as well as performance. Many of the companies represented at the standards group meetings have significant vested interests in the outcome of the standardisation process. In some cases, the company behind a particular proposal may own key patents that could lead to lucrative licensing revenue if the proposal is adopted as part of a worldwide standard. The JVT has asked members to identify any patented technology in proposals and is attempting to ensure that the Baseline Profile of H.264 can be implemented without royalty payments to patent holders, in order to promote widespread adoption of the standard.

On the practical side, the standards bodies work to tight timescales and it may be necessary to compromise on some aspects of the work plan in order to complete the standard on time, since it is important to ensure that the standard is technically competitive when it is released. In some ways the MPEG-4 Visual and H.264 standardisation efforts have been a race against time, since many proprietary (nonstandard) image and video coding systems and packages continue to be developed and released. Outside the constraints of the standardisation process, it is possible for a proprietary system to gain market share quickly and to make a standard effectively obsolete by the time it is published.

4.3 USING THE STANDARDS

MPEG-4 Visual (Second Edition) is a document consisting of 539 A4 pages and H.264 is over 250 A4 pages long. The standards are specified in detail in order to attempt to ensure compatibility between conforming systems and products but the level of detail in the standards documents can make it difficult to extract information from them. In this section we give an overview of the content of the standards and suggest how to interpret and apply them.

	Object types														
Profile	Simple	Advanced Simple	Advanced Real-Time Simple	Core	Main	Advanced Coding Efficiency	N-bit	Simple Scalable	Fine Granular Scalability	Core Scalable	Scalable Texture	Advanced Scalable Texture	Simple Studio	Core Studio	Simple Face Animation
Simple	✓														
Advanced Simple	✓	✓													
Advanced Real-Time Simple	✓		✓												
Core	✓			✓											
Advanced Core	✓			✓							✓				
Main	✓			✓	✓						✓				
Advanced Coding Efficiency	✓			✓		✓									
N-bit	✓			✓			✓								
Simple Scalable	✓							✓							
Fine Granular Scalability	✓	✓							✓						
Core Scalable	✓			✓				✓		✓					
Scalable Texture											✓				
Advanced Scalable Texture												✓			
Simple Studio													✓		
Core Studio													✓	✓	
Basic Animated Texture											✓				✓
Simple Face Animation														✓	
Simple FBA															✓
Hybrid	✓			✓							✓			✓	✓

Figure 5.1 MPEG-4 Visual profiles and objects

Figure 5.1 lists each of the MPEG-4 Visual profiles (left-hand column) and visual object types (top row). The table entries indicate which object types are contained within each profile. For example, a CODEC compatible with Simple Profile must be capable of coding and decoding Simple objects and a Core Profile CODEC must be capable of coding and decoding Simple *and* Core objects.

Profiles are an important mechanism for encouraging interoperability between CODECs from different manufacturers. The MPEG-4 Visual standard describes a diverse range of coding tools and it is unlikely that any commercial CODEC would require the implementation of all the tools. Instead, a CODEC designer chooses a profile that contains adequate tools for the target application. For example, a basic CODEC implemented on a low-power processor may use Simple profile, a CODEC for streaming video applications may choose Advanced Real Time Simple and so on. To date, some profiles have had more of an impact on the marketplace than others. The Simple and Advanced Simple profiles are particularly popular with manufacturers and users whereas the profiles for the coding of arbitrary-shaped objects have had very limited commercial impact (see Chapter 8 for further discussion of the commercial impact of MPEG-4 Profiles).

Profiles define a subset of coding tools and *Levels* define constraints on the parameters of the bitstream. Table 5.3 lists the Levels for the popular Simple-based profiles (Simple,

Table 5.3 Levels for Simple-based profiles

Profile	Level	Typical resolution	Max. bitrate	Max. objects
Simple	L0	176 × 144	64 kbps	1 simple
	L1	176 × 144	64 kbps	4 simple
	L2	352 × 288	128 kbps	4 simple
	L3	352 × 288	384 kbps	4 simple
Advanced Simple (AS)	L0	176 × 144	128 kbps	1 AS or simple
	L1	176 × 144	128 kbps	4 AS or simple
	L2	352 × 288	384 kbps	4 AS or simple
	L3	352 × 288	768 kbps	4 AS or simple
	L4	352 × 576	3 Mbps	4 AS or simple
	L5	720 × 576	8 Mbps	4 AS or simple
Advanced Real-Time Simple (ARTS)	L1	176 × 144	64 kbps	4 ARTS or simple
	L2	352 × 288	128 kbps	4 ARTS or simple
	L3	352 × 288	384 kbps	4 ARTS or simple
	L4	352 × 288	2 Mbps	16 ARTS or simple

Advanced Simple and Advanced Real Time Simple). Each Level places constraints on the maximum performance required to decode an MPEG-4 coded sequence. For example, a multimedia terminal with limited processing capabilities and a small amount of memory may only support Simple Profile @ Level 0 bitstream decoding. The Level definitions place restrictions on the amount of buffer memory, the decoded frame size and processing rate (in macroblocks per second) and the number of video objects (one in this case, a single rectangular frame). A terminal that can cope with these parameters is guaranteed to be capable of successfully decoding any conforming Simple Profile @ Level 0 bitstream. Higher Levels of Simple Profile require a decoder to handle up to four Simple Profile video objects (for example, up to four rectangular objects covering the QCIF or CIF display resolution).

5.2.3 Video Objects

One of the key contributions of MPEG-4 Visual is a move away from the ‘traditional’ view of a video sequence as being merely a collection of rectangular frames of video. Instead, MPEG-4 Visual treats a video sequence as a collection of one or more *video objects*. MPEG-4 Visual defines a video object as a flexible ‘entity that a user is allowed to access (seek, browse) and manipulate (cut and paste)’ [1]. A video object (VO) is an area of the video scene that may occupy an arbitrarily-shaped region and may exist for an arbitrary length of time. An instance of a VO at a particular point in time is a *video object plane* (VOP).

This definition encompasses the traditional approach of coding complete frames, in which each VOP is a single frame of video and a sequence of frames forms a VO (for example, Figure 5.2 shows a VO consisting of three rectangular VOPs). However, the introduction of the VO concept allows more flexible options for coding video. Figure 5.3 shows a VO that consists of three irregular-shaped VOPs, each one existing within a frame and each coded separately (*object-based coding*).

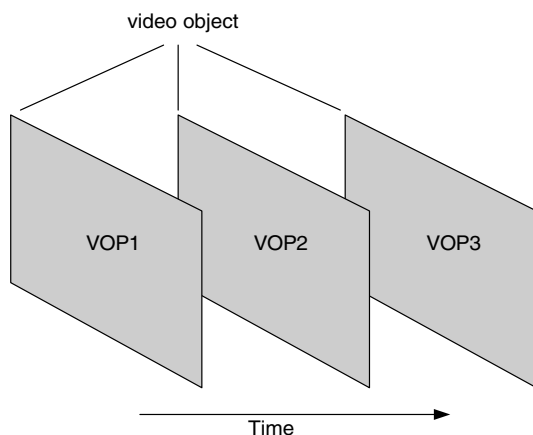


Figure 5.2 VOPs and VO (rectangular)

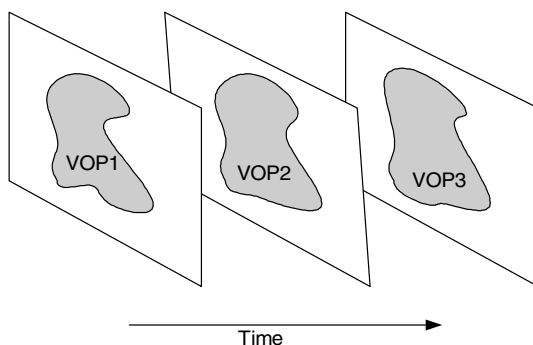


Figure 5.3 VOPs and VO (arbitrary shape)

A video scene (e.g. Figure 5.4) may be made up of a background object (VO3 in this example) and a number of separate foreground objects (VO1, VO2). This approach is potentially much more flexible than the fixed, rectangular frame structure of earlier standards. The separate objects may be coded with different visual qualities and temporal resolutions to reflect their ‘importance’ to the final scene, objects from multiple sources (including synthetic and ‘natural’ objects) may be combined in a single scene and the composition and behaviour of the scene may be manipulated by an end-user in highly interactive applications. Figure 5.5 shows a new video scene formed by adding VO1 from Figure 5.4, a new VO2 and a new background VO. Each object is coded separately using MPEG-4 Visual (the compositing of visual and audio objects is assumed to be handled separately, for example by MPEG-4 Systems [2]).

5.3 CODING RECTANGULAR FRAMES

Notwithstanding the potential flexibility offered by object-based coding, the most popular application of MPEG-4 Visual is to encode complete frames of video. The tools required

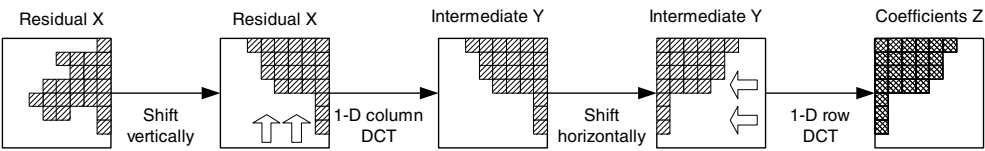


Figure 5.55 Shape-adaptive DCT

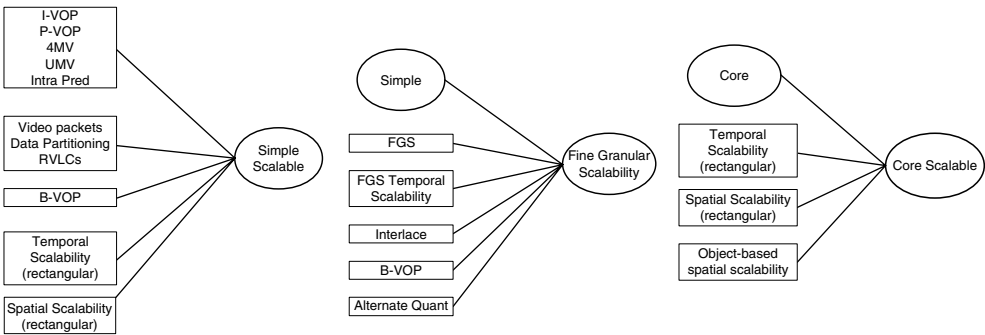


Figure 5.56 Tools and objects for scalable coding

contain one or more transparent pixels. The Forward SA-DCT consists of the following steps (Figure 5.55):

- 1. Shift opaque residual values *X* to the top of the 8×8 block.
- 2. Apply a 1D DCT to each column (the number of points in the transform matches the number of opaque values in each column).
- 3. Shift the resulting intermediate coefficients *Y* to the left of the block.
- 4. Apply a 1D DCT to each row (matched to the number of values in each row).

The final coefficients (*Z*) are quantised, zigzag scanned and encoded. The decoder reverses the process (making use of the shape information decoded from the BAB) to reconstruct the 8×8 block of samples. The SA-DCT is more complex than the normal 8×8 DCT but can improve coding efficiency for boundary MBs.

5.4.4 The N-bit Profile

The N-bit profile contains Simple and Core objects plus the *N-bit* tool. This supports coding of luminance and chrominance data containing between four and twelve bits per sample (instead of the usual restriction to eight bits per sample). Possible applications of the N-bit profile include video coding for displays with low colour depth (where the limited display capability means that less than eight bits are required to represent each sample) or for high-quality display applications (where the display has a colour depth of more than eight bits per sample and high coded fidelity is desired).

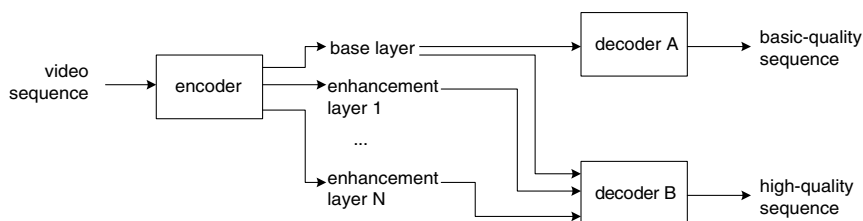


Figure 5.57 Scalable coding: general concept

5.5 SCALABLE VIDEO CODING

Scalable encoding of video data enables a decoder to decode selectively only part of the coded bitstream. The coded stream is arranged in a number of *layers*, including a ‘base’ layer and one or more ‘enhancement’ layers (Figure 5.57). In this figure, decoder A receives only the base layer and can decode a ‘basic’ quality version of the video scene, whereas decoder B receives all layers and decodes a high quality version of the scene. This has a number of applications, for example, a low-complexity decoder may only be capable of decoding the base layer; a low-rate bitstream may be extracted for transmission over a network segment with limited capacity; and an error-sensitive base layer may be transmitted with higher priority than enhancement layers.

MPEG-4 Visual supports a number of scalable coding modes. *Spatial scalability* enables a (rectangular) VOP to be coded at a hierarchy of spatial resolutions. Decoding the base layer produces a low-resolution version of the VOP and decoding successive enhancement layers produces a progressively higher-resolution image. *Temporal scalability* provides a low frame-rate base layer and enhancement layer(s) that build up to a higher frame rate. The standard also supports *quality scalability*, in which the enhancement layers improve the visual quality of the VOP and *complexity scalability*, in which the successive layers are progressively more complex to decode. *Fine Grain Scalability* (FGS) enables the quality of the sequence to be increased in small steps. An application for FGS is streaming video across a network connection, in which it may be useful to scale the coded video stream to match the available bit rate as closely as possible.

5.5.1 Spatial Scalability

The base layer contains a reduced-resolution version of each coded frame. Decoding the base layer alone produces a low-resolution output sequence and decoding the base layer with enhancement layer(s) produces a higher-resolution output. The following steps are required to *encode* a video sequence into two spatial layers:

1. Subsample each input video frame (Figure 5.58) (or video object) horizontally and vertically (Figure 5.59).
2. Encode the reduced-resolution frame to form the base layer.
3. Decode the base layer and up-sample to the original resolution to form a prediction frame (Figure 5.60).
4. Subtract the full-resolution frame from this prediction frame (Figure 5.61).
5. Encode the difference (residual) to form the enhancement layer.



Figure 5.58 Original video frame



Figure 5.59 Sub-sampled frame to be encoded as base layer



Figure 5.60 Base layer frame (decoded and upsampled)



Figure 5.61 Residual to be encoded as enhancement layer

A single-layer decoder decodes only the base layer to produce a reduced-resolution output sequence. A two-layer decoder can reconstruct a full-resolution sequence as follows:

1. Decode the base layer and up-sample to the original resolution.
2. Decode the enhancement layer.
3. Add the decoded residual from the enhancement layer to the decoded base layer to form the output frame.

An **I-VOP** in an enhancement layer is encoded without any spatial prediction, i.e. as a complete frame or object at the enhancement resolution. In an enhancement layer **P-VOP**, the decoded, up-sampled base layer VOP (at the same position in time) is used as a prediction *without* any motion compensation. The difference between this prediction and the input frame is encoded using the texture coding tools, i.e. no motion vectors are transmitted for an enhancement P-VOP. An enhancement layer **B-VOP** is predicted from two directions. The backward prediction is formed by the decoded, up-sampled base layer VOP (at the same position in time), without any motion compensation (and hence without any MVs). The forward prediction is formed by the previous VOP in the enhancement layer (even if this is itself a B-VOP), with motion-compensated prediction (and hence MVs).

If the VOP has arbitrary (binary) shape, a base layer and enhancement layer BAB is required for each MB. The base layer BAB is encoded as usual, based on the shape and size of the base layer object. A BAB in a P-VOP enhancement layer is coded using prediction from an up-sampled version of the base layer BAB. A BAB in a B-VOP enhancement layer may be coded in the same way, or using forward prediction from the previous enhancement VOP (as described in Section 5.4.1.1).

5.5.2 Temporal Scalability

The base layer of a temporal scalable sequence is encoded at a low video frame rate and a temporal enhancement layer consists of I-, P- and/or B-VOPs that can be decoded together with the base layer to provide an increased video frame rate. Enhancement layer VOPs are predicted using motion-compensated prediction according to the following rules.

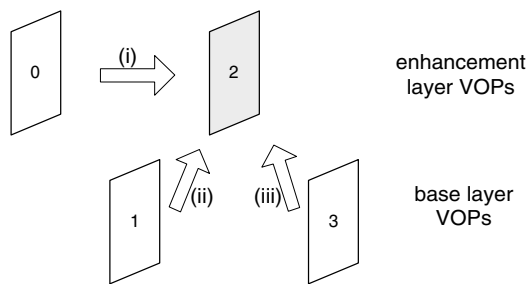


Figure 5.62 Temporal enhancement P-VOP prediction options

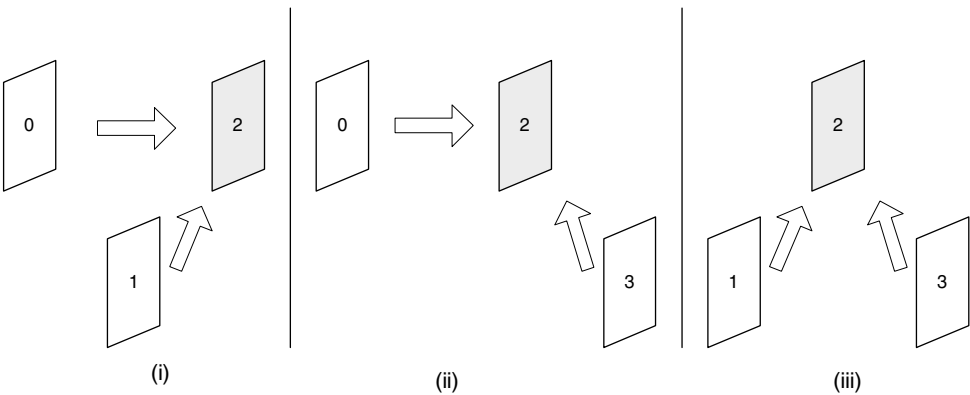


Figure 5.63 Temporal enhancement B-VOP prediction options

An enhancement *I-VOP* is encoded without any prediction. An enhancement *P-VOP* is predicted from (i) the previous enhancement VOP, (ii) the previous base layer VOP or (iii) the next base layer VOP (Figure 5.62). An enhancement *B-VOP* is predicted from (i) the previous enhancement and previous base layer VOPs, (ii) the previous enhancement and next base layer VOPs or (iii) the previous and next base layer VOPs (Figure 5.63).

5.5.3 Fine Granular Scalability

Fine Granular Scalability (FGS) [5] is a method of encoding a sequence as a base layer and enhancement layer. The enhancement layer can be truncated during or after encoding (reducing the bitrate and the decoded quality) to give highly flexible control over the transmitted bitrate. FGS may be useful for video streaming applications, in which the available transmission bandwidth may not be known in advance. In a typical scenario, a sequence is coded as a base layer and a high-quality enhancement layer. Upon receiving a request to send the sequence at a particular bitrate, the streaming server transmits the base layer and a truncated version of the enhancement layer. The amount of truncation is chosen to match the available transmission bitrate, hence maximising the quality of the decoded sequence without the need to re-encode the video clip.

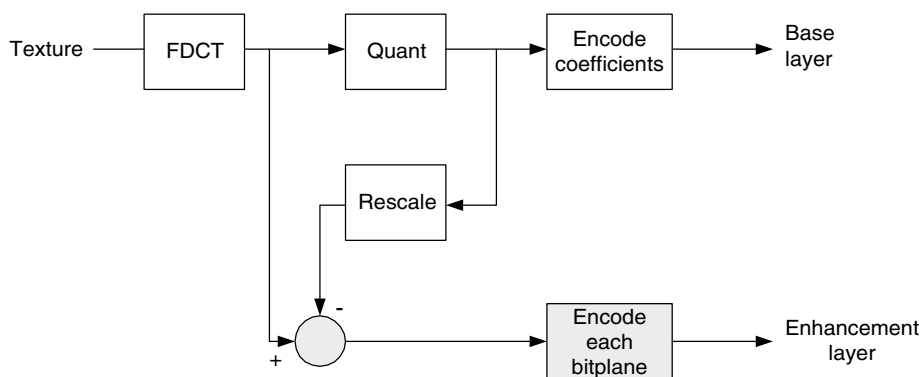


Figure 5.64 FGS encoder block diagram (simplified)

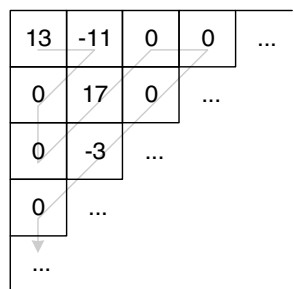


Figure 5.65 Block of residual coefficients (top-left corner)

Encoding

Figure 5.64 shows a simplified block diagram of an FGS encoder (motion compensation is not shown). In the Base Layer, the texture (after motion compensation) is transformed with the forward DCT, quantised and encoded. The quantised coefficients are re-scaled (‘inverse quantised’) and these re-scaled coefficients are subtracted from the unquantised DCT coefficients to give a set of difference coefficients. The difference coefficients for each block are encoded as a series of *bitplanes*. First, the residual coefficients are reordered using a zigzag scan. The highest-order bits of each coefficient (zeros or ones) are encoded first (the MS bit-plane) followed by the next highest-order bits and so on until the LS bits have been encoded.

Example

A block of residual coefficients is shown in Figure 5.65 (coefficients not shown are zero). The coefficients are reordered in a zigzag scan to produce the following list:

$$+13, -11, 0, 0, +17, 0, 0, 0, -3, 0, 0 \dots$$

The bitplanes corresponding to the magnitude of each residual coefficient are shown in Table 5.6. In this case, the highest plane containing nonzero bits is plane 4 (because the highest magnitude is 17).