DIGITAL VIDEO PROCESSING

Second Edition
To Sevim and Kaya Tekalp, my mom and dad,
To Özge, my beloved wife, and
To Engin Deniz, my son, and Derya Cansu, my daughter
This page intentionally left blank
Contents

Preface xvii
About the Author xxv

1 Multi-Dimensional Signals and Systems 1

1.1 Multi-Dimensional Signals 2
  1.1.1 Finite-Extent Signals and Periodic Signals 2
  1.1.2 Symmetric Signals 5
  1.1.3 Special Multi-Dimensional Signals 5

1.2 Multi-Dimensional Transforms 8
  1.2.1 Fourier Transform of Continuous Signals 8
  1.2.2 Fourier Transform of Discrete Signals 12
  1.2.3 Discrete Fourier Transform (DFT) 14
  1.2.4 Discrete Cosine Transform (DCT) 18

1.3 Multi-Dimensional Systems 20
  1.3.1 Impulse Response and 2D Convolution 20
  1.3.2 Frequency Response 23
  1.3.3 FIR Filters and Symmetry 25
  1.3.4 IIR Filters and Partial Difference Equations 27

1.4 Multi-Dimensional Sampling Theory 30
  1.4.1 Sampling on a Lattice 30
  1.4.2 Spectrum of Signals Sampled on a Lattice 34
  1.4.3 Nyquist Criterion for Sampling on a Lattice 36
1.4.4 Reconstruction from Samples on a Lattice 41
1.5 Sampling Structure Conversion 42
References 47
Exercises 48
Problem Set 1 48
MATLAB Exercises 50

2 Digital Images and Video 53
2.1 Human Visual System and Color 54
2.1.1 Color Vision and Models 54
2.1.2 Contrast Sensitivity 57
2.1.3 Spatio-Temporal Frequency Response 59
2.1.4 Stereo/Depth Perception 62
2.2 Analog Video 63
2.2.1 Progressive vs. Interlaced Scanning 64
2.2.2 Analog-Video Signal Formats 65
2.2.3 Analog-to-Digital Conversion 66
2.3 Digital Video 67
2.3.1 Spatial Resolution and Frame Rate 67
2.3.2 Color, Dynamic Range, and Bit-Depth 69
2.3.3 Color Image Processing 71
2.3.4 Digital-Video Standards 74
2.4 3D Video 79
2.4.1 3D-Display Technologies 79
2.4.2 Stereoscopic Video 82
2.4.3 Multi-View Video 83
2.5 Digital-Video Applications 85
2.5.1 Digital TV 85
2.5.2 Digital Cinema 89
2.5.3 Video Streaming over the Internet 92
2.5.4 Computer Vision and Scene/Activity Understanding 95
2.6 Image and Video Quality 96
2.6.1 Visual Artifacts 96
2.6.2 Subjective Quality Assessment 97
2.6.3 Objective Quality Assessment 98
References 100
3 Image Filtering 105
3.1 Image Smoothing 106
  3.1.1 Linear Shift-Invariant Low-Pass Filtering 106
  3.1.2 Bi-Lateral Filtering 109
3.2 Image Re-Sampling and Multi-Resolution Representations 110
  3.2.1 Image Decimation 111
  3.2.2 Interpolation 113
  3.2.3 Multi-Resolution Pyramid Representations 120
  3.2.4 Wavelet Representations 121
3.3 Image-Gradient Estimation, Edge and Feature Detection 127
  3.3.1 Estimation of the Image Gradient 128
  3.3.2 Estimation of the Laplacian 132
  3.3.3 Canny Edge Detection 134
  3.3.4 Harris Corner Detection 135
3.4 Image Enhancement 137
  3.4.1 Pixel-Based Contrast Enhancement 137
  3.4.2 Spatial Filtering for Tone Mapping and Image Sharpening 142
3.5 Image Denoising 147
  3.5.1 Image and Noise Models 148
  3.5.2 Linear Space-Invariant Filters in the DFT Domain 150
  3.5.3 Local Adaptive Filtering 153
  3.5.4 Nonlinear Filtering: Order-Statistics, Wavelet Shrinkage, and Bi-Lateral Filtering 158
  3.5.5 Non-Local Filtering: NL-Means and BM3D 162
3.6 Image Restoration 164
  3.6.1 Blur Models 165
  3.6.2 Restoration of Images Degraded by Linear Space-Invariant Blurs 169
  3.6.3 Blind Restoration – Blur Identification 175
  3.6.4 Restoration of Images Degraded by Space-Varying Blurs 177
  3.6.5 Image In-Painting 180
References 181
Exercises 186
  Problem Set 3 186
  MATLAB Exercises 189
MATLAB Resources 193
4 Motion Estimation 195

4.1 Image Formation 196
   4.1.1 Camera Models 196
   4.1.2 Photometric Effects of 3D Motion 201

4.2 Motion Models 202
   4.2.1 Projected Motion vs. Apparent Motion 203
   4.2.2 Projected 3D Rigid-Motion Models 207
   4.2.3 2D Apparent-Motion Models 210

4.3 2D Apparent-Motion Estimation 214
   4.3.1 Sparse Correspondence, Optical-Flow Estimation, and Image-Registration Problems 214
   4.3.2 Optical-Flow Equation and Normal Flow 217
   4.3.3 Displaced-Frame Difference 219
   4.3.4 Motion Estimation is Ill-Posed: Occlusion and Aperture Problems 220
   4.3.5 Hierarchical Motion Estimation 223
   4.3.6 Performance Measures for Motion Estimation 224

4.4 Differential Methods 225
   4.4.1 Lukas–Kanade Method 225
   4.4.2 Horn–Schunk Motion Estimation 230

4.5 Matching Methods 233
   4.5.1 Basic Block-Matching 234
   4.5.2 Variable-Size Block-Matching 238
   4.5.3 Hierarchical Block-Matching 240
   4.5.4 Generalized Block-Matching – Local Deformable Motion 241
   4.5.5 Homography Estimation from Feature Correspondences 243

4.6 Nonlinear Optimization Methods 245
   4.6.1 Pel-Recursive Motion Estimation 245
   4.6.2 Bayesian Motion Estimation 247

4.7 Transform-Domain Methods 249
   4.7.1 Phase-Correlation Method 249
   4.7.2 Space-Frequency Spectral Methods 251

4.8 3D Motion and Structure Estimation 251
   4.8.1 Camera Calibration 252
   4.8.2 Affine Reconstruction 253
   4.8.3 Projective Reconstruction 255
   4.8.4 Euclidean Reconstruction 260
5 Video Segmentation and Tracking 273

5.1 Image Segmentation 275
   5.1.1 Thresholding 275
   5.1.2 Clustering 277
   5.1.3 Bayesian Methods 281
   5.1.4 Graph-Based Methods 285
   5.1.5 Active-Contour Models 287

5.2 Change Detection 289
   5.2.1 Shot-Boundary Detection 289
   5.2.2 Background Subtraction 291

5.3 Motion Segmentation 298
   5.3.1 Dominant-Motion Segmentation 299
   5.3.2 Multiple-Motion Segmentation 302
   5.3.3 Region-Based Motion Segmentation: Fusion of Color and Motion 311
   5.3.4 Simultaneous Motion Estimation and Segmentation 313

5.4 Motion Tracking 317
   5.4.1 Graph-Based Spatio-Temporal Segmentation and Tracking 319
   5.4.2 Kanade–Lucas–Tomasi Tracking 319
   5.4.3 Mean-Shift Tracking 321
   5.4.4 Particle-Filter Tracking 323
   5.4.5 Active-Contour Tracking 325
   5.4.6 2D-Mesh Tracking 327

5.5 Image and Video Matting 328

5.6 Performance Evaluation 330
References 331
MATLAB Exercises 338
Internet Resources 339
6 Video Filtering  341
   6.1 Theory of Spatio-Temporal Filtering  342
       6.1.1 Frequency Spectrum of Video  342
       6.1.2 Motion-Adaptive Filtering  345
       6.1.3 Motion-Compensated Filtering  345
   6.2 Video-Format Conversion  349
       6.2.1 Down-Conversion  351
       6.2.2 De-Interlacing  355
       6.2.3 Frame-Rate Conversion  361
   6.3 Multi-Frame Noise Filtering  367
       6.3.1 Motion-Adaptive Noise Filtering  367
       6.3.2 Motion-Compensated Noise Filtering  369
   6.4 Multi-Frame Restoration  374
       6.4.1 Multi-Frame Modeling  375
       6.4.2 Multi-Frame Wiener Restoration  375
   6.5 Multi-Frame Super-Resolution  377
       6.5.1 What Is Super-Resolution?  378
       6.5.2 Modeling Low-Resolution Sampling  381
       6.5.3 Super-Resolution in the Frequency Domain  386
       6.5.4 Multi-Frame Spatial-Domain Methods  389
References  394
Exercises  399
   Problem Set 6  399
   MATLAB Exercises  400

7 Image Compression  401
   7.1 Basics of Image Compression  402
       7.1.1 Information Theoretic Concepts  402
       7.1.2 Elements of Image-Compression Systems  405
       7.1.3 Quantization  406
       7.1.4 Symbol Coding  409
       7.1.5 Huffman Coding  410
       7.1.6 Arithmetic Coding  414
   7.2 Lossless Image Compression  417
       7.2.1 Bit-Plane Coding  418
       7.2.2 Run-Length Coding and ITU G3/G4 Standards  419
       7.2.3 Adaptive Arithmetic Coding and JBIG  423
8 Video Compression 461

8.1 Video-Compression Approaches 462
  8.1.1 Intra-Frame Compression, Motion JPEG 2000, and Digital Cinema 462
  8.1.2 3D-Transform Coding 463
  8.1.3 Motion-Compensated Transform Coding 466

8.2 Early Video-Compression Standards 467
  8.2.1 ISO and ITU Standards 467
  8.2.2 MPEG-1 Standard 468
  8.2.3 MPEG-2 Standard 476

8.3 MPEG-4 AVC/ITU-T H.264 Standard 483
  8.3.1 Input-Video Formats and Data Structure 484
  8.3.2 Intra-Prediction 485
  8.3.3 Motion Compensation 486
  8.3.4 Transform 488
  8.3.5 Other Tools and Improvements 489

8.4 High-Efficiency Video-Coding (HEVC) Standard 491
  8.4.1 Video-Input Format and Data Structure 491
  8.4.2 Coding-Tree Units 492
  8.4.3 Tools for Parallel Encoding/Decoding 493
  8.4.4 Other Tools and Improvements 495

8.5 Scalable-Video Compression 497
  8.5.1 Temporal Scalability 498
## Contents

8.5.2  Spatial Scalability  499  
8.5.3  Quality (SNR) Scalability  500  
8.5.4  Hybrid Scalability  502  

8.6  Stereo and Multi-View Video Compression  502  
8.6.1  Frame-Compatible Stereo-Video Compression  503  
8.6.2  Stereo and Multi-View Video-Coding Extensions of the H.264/AVC Standard  504  
8.6.3  Multi-View Video Plus Depth Compression  507  

References  512  
Exercises  514  
Internet Resources  515

### A  Vector-Matrix Operations in Image and Video Processing  517  
A.1  Two-Dimensional Convolution  517  
A.2  Two-Dimensional Discrete-Fourier Transform  520  
  A.2.1  Diagonalization of Block-Circulant Matrices  521  
A.3  Three-Dimensional Rotation – Rotation Matrix  521  
  A.3.1  Euler Angles  522  
  A.3.2  Rotation About an Arbitrary Axis  523  
  A.3.3  Quaternion Representation  524  

References  525  
Exercises  525

### B  Ill-Posed Problems in Image and Video Processing  527  
B.1  Image Representations  527  
  B.1.1  Deterministic Framework – Function/Vector Spaces  527  
  B.1.2  Bayesian Framework – Random Fields  528  
B.2  Overview of Image Models  528  
B.3  Basics of Sparse-Image Modeling  530  
B.4  Well-Posed Formulations of Ill-Posed Problems  531  
  B.4.1  Constrained-Optimization Problem  531  
  B.4.2  Bayesian-Estimation Problem  532  

References  532

### C  Markov and Gibbs Random Fields  533  
C.1  Equivalence of Markov Random Fields and Gibbs Random Fields  533  
  C.1.1  Markov Random Fields  534  
  C.1.2  Gibbs Random Fields  535  
  C.1.3  Equivalence of MRF and GRF  536
Contents

C.2 Gibbs Distribution as an *a priori* PDF Model 537
C.3 Computation of Local Conditional Probabilities from a Gibbs Distribution 538
References 539

D Optimization Methods 541
D.1 Gradient-Based Optimization 542
  D.1.1 Steepest-Descent Method 542
  D.1.2 Newton–Raphson Method 543
D.2 Simulated Annealing 544
  D.2.1 Metropolis Algorithm 545
  D.2.2 Gibbs Sampler 546
D.3 Greedy Methods 547
  D.3.1 Iterated Conditional Modes 547
  D.3.2 Mean-Field Annealing 548
  D.3.3 Highest Confidence First 548
References 549

E Model Fitting 551
E.1 Least-Squares Fitting 551
E.2 Least-Squares Solution of Homogeneous Linear Equations 552
  E.2.1 Alternate Derivation 553
E.3 Total Least-Squares Fitting 554
E.4 Random-Sample Consensus (RANSAC) 556
References 556

Index 557
This page intentionally left blank
Preface

The first edition of this book (1995) was the first comprehensive textbook on digital video processing. However, digital video technologies and video processing algorithms were not mature enough then. Digital TV standards were just being written, digital cinema was not even in consideration, and digital video cameras and DVD were just entering the market. Hence, the first edition contained some now-outdated methods/algorithms and technologies compared with the state of the art today, and obviously missed important developments in the last 20 years. The first edition was organized into 25 smaller chapters on what were then conceived to be important topics in video processing, each intended to be covered in one or two lectures during a one-semester course. Some methods covered in the first edition—e.g., pel-recursive motion estimation, vector quantization, fractal compression, and model-based coding—no longer reflect the state of the art. Some technologies covered in the first edition, such as analog video/TV and 128K videophone, are now obsolete.

In the 20 years since the first edition, digital video has become ubiquitous in our daily lives in the digital age. Video processing algorithms have become more mature with significant new advances made by signal processing and computer vision communities, and the most popular and successful techniques and algorithms for different tasks have become clearer. Hence, it is now the right time for an updated edition of the book. This book aims to fill the need for a comprehensive, rigorous, and tutorial-style textbook for digital image and video processing that covers the most recent state of the art in a well-balanced manner.

This second edition significantly improves the organization of the material and presentation style and updates the technical content with the most up-to-date techniques, successful algorithms, and most recent knowledge in the field. It is
organized into eight comprehensive chapters, where each covers a major subject, including multi-dimensional signal processing, image/video basics, image filtering, motion estimation, video segmentation, video filtering, image compression, and video compression, with an emphasis on the most successful techniques in each subject area. Therefore, this is not an incremental revision—it is almost a complete rewrite.

The book is intended as a quantitative textbook for advanced undergraduate- and graduate-level classes on digital image and video processing. It assumes familiarity with calculus, linear algebra, probability, and some basic digital signal processing concepts. Readers with a computer science background who may not be familiar with the fundamental signal processing concepts can skip Chapter 1 and still follow the remaining chapters reasonably well. Although the presentation is rigorous, it is in a tutorial style starting from fundamentals. Hence, it can also be used as a reference book or for self-study by researchers and engineers in the industry or in academia. This book enables the reader to

- understand theoretical foundations of image and video processing methods,
- learn the most popular and successful algorithms to solve common image and video processing problems,
- reinforce their understanding by solving problems at the end of each chapter, and
- practice methods by doing the MATLAB projects at the end of each chapter.

Digital video processing refers to manipulation of the digital video bitstream. All digital video applications require compression. In addition, they may benefit from filtering for format conversion, enhancement, restoration, and super-resolution in order to obtain better-quality images or to extract specific information, and some may require additional processing for motion estimation, video segmentation, and 3D scene analysis. What makes digital video processing different from still image processing is that video contains a significant amount of temporal correlation (redundancy) between the frames. One may attempt to process video as a sequence of still images, where each frame is processed independently. However, multi-frame processing techniques using inter-frame correlations enable us to develop more effective algorithms, such as motion-compensated filtering and prediction. In addition, some tasks, such as motion estimation or the analysis of a time-varying scene, obviously cannot be performed on the basis of a single image.

It is the goal of this book to provide the reader with the mathematical basis of image (single-frame) and video (multi-frame) processing methods. In particular, this book answers the following fundamental questions:
• How do we separate images (signal) from noise?
• Is there a relationship between interpolation, restoration, and super-resolution?
• How do we estimate 2D and 3D motion for different applications?
• How do we segment images and video into regions of interest?
• How do we track objects in video?
• Is video filtering a better-posed problem than image filtering?
• What makes super-resolution possible?
• Can we obtain a high-quality still image from a video clip?
• What makes image and video compression possible?
• How do we compress images and video?
• What are the most recent international standards for image/video compression?
• What are the most recent standards for 3D video representation and compression?

Most image and video processing problems are ill-posed (underdetermined and/or sensitive to noise) and their solutions rely on some sort of image and video models. Approaches to *image modeling* for solution of ill-posed problems are discussed in Appendix B. In particular, image models can be classified as those based on

- local smoothness,
- sparseness in a transform domain, and
- non-local self-similarity.

Most image processing algorithms employ one or more of these models. Video models use, in addition to the above,

- global or block translation motion,
- parametric motion,
- motion (spatial) smoothness,
- motion uniformity in time (temporal continuity or smoothness), and
- planar support in 3D spatio-temporal frequency domain.

An overview of the chapters follows.

Chapter 1 reviews the basics of multi-dimensional signals, transforms, and systems, which form the theoretical basis of many image and video processing methods. We also address spatio-temporal sampling on MD lattices, which includes several practical sampling structures such as progressive and interlaced sampling, as well as theory of sampling structure conversion. Readers with a computer science background who may not be familiar with signal processing concepts can skip this chapter and start with Chapter 2.
Chapter 2 aims to provide a basic understanding of digital image and video fundamentals. We cover the basic concepts of human vision, spatial frequency, color models, analog and digital video representations, digital video standards, 3D stereo and multi-view video representations, and evaluation of digital video quality. We introduce popular digital video applications, including digital TV, digital cinema, and video streaming over the Internet.

Chapter 3 addresses image (still-frame) filtering problems such as image resampling (decimation and interpolation), gradient estimation and edge detection, enhancement, de-noising, and restoration. Linear shift-invariant, adaptive, and non-linear filters are considered. We provide a general framework for solution of ill-posed inverse problems in Appendix B.

Chapter 4 covers 2D and 3D motion estimation methods. Motion estimation is at the heart of digital video processing since motion is the most prominent feature of video, and motion-compensated filtering is the most effective way to utilize temporal redundancy. Furthermore, many computer vision tasks require 2D or 3D motion estimation and tracking as a first step. 2D motion estimation, which refers to dense optical flow or sparse feature correspondence estimation, can be based on nonparametric or parametric methods. Nonparametric methods include image gradient-based optical flow estimation, block matching, pel-recursive methods, Bayesian methods, and phase correlation methods. The parametric methods, based on the affine model or the homography, can be used for image registration or to estimate local deformations. 3D motion/structure estimation methods include those based on the two-frame epipolar constraint (mainly for stereo pairs) or multi-frame factorization methods. Reconstruction of Euclidean 3D structure requires full-camera calibration while projective reconstruction can be performed without any calibration.

Chapter 5 introduces image segmentation and change detection, as well as segmentation of dominant motion or multiple motions using parameter clustering and Bayesian methods. We also discuss simultaneous motion estimation and segmentation. Since two-view motion estimation techniques are very sensitive to inaccuracies in the estimates of image gradients or point correspondences, motion tracking of segmented objects over long monocular sequences or stereo pairs, which yield more robust results, are also considered.

Chapter 6 addresses video filtering, including standards conversion, de-noising, and super-resolution. It starts with the basic theory of motion-compensated filtering. Next, standards conversion problems, including frame rate conversion and de-interlacing, are covered. Video frames often suffer from graininess, especially when viewed in freeze-frame mode. Hence, motion-adaptive and motion-compensated
filtering for noise suppression are discussed. Finally, a comprehensive model for low-resolution video acquisition and super-resolution reconstruction methods (based on this model) that unify various video filtering problems are presented.

Chapter 7 covers still-image, including binary (FAX) and gray-scale image, compression methods and standards such as JPEG and JPEG 2000. In particular, we discuss lossless image compression and lossy discrete cosine transform coding and wavelet coding methods.

Chapter 8 discusses video compression methods and standards that have made digital video applications such as digital TV and digital cinema a reality. After a brief introduction to different approaches to video compression, we cover MPEG-2, AVC/H.264, and HEVC standards in detail, as well as their scalable video coding and stereo/multi-view video coding extensions.

This textbook is the outcome of my experience in teaching digital image and video processing for more than 20 years. It is comprehensive, written in a tutorial style, which covers both fundamentals and the most recent progress in image filtering, motion estimation and tracking, image/video segmentation, video filtering, and image/video compression with equal emphasis on these subjects. Unfortunately, it is not possible to cover all state-of-the-art methods in digital video processing and computer vision in a tutorial style in a single volume. Hence, only the most fundamental, popular techniques and algorithms are explained in a tutorial style. More advanced algorithms and recent research results are briefly summarized and references are provided for self-study. Problem sets and MATLAB projects are included at the end of each chapter for the reader to practice the methods.

Teaching materials will be provided to instructors upon request. A teaching plan is provided in Table P.1, which assumes a 14-week semester with two 75-minute classes each week, to cover the whole book in a one-semester digital image and video processing course. Alternatively, it is possible to cover the book in two semesters, which would allow time to delve into more technical details with each subject. The first semester can be devoted to digital image processing, covering Chapters 1, 2, 3, and 7. In the second semester, Chapters 4, 5, 6, and 8 can be covered in a follow-up digital video processing course.

Clearly, this book is a compilation of knowledge collectively created by the signal processing and computer science communities. I have included many citations and references in each chapter, but I am sure I have neglected some since it is impossible to give credit to all outstanding academic and industrial researchers who contributed to the development of image and video processing. Furthermore, outstanding innovations in image and video coding are a result of work done by many scientists
### Table P.1  Suggested Teaching Plan for a One-Semester Course

<table>
<thead>
<tr>
<th>Lecture</th>
<th>Topic</th>
<th>Chapter/Sections</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2D signals, 2D transforms</td>
<td>1.1, 1.2</td>
</tr>
<tr>
<td>2</td>
<td>2D systems, 2D FIR filters, frequency response</td>
<td>1.3</td>
</tr>
<tr>
<td>3</td>
<td>MD spatio-temporal sampling on lattices</td>
<td>1.4, 1.5</td>
</tr>
<tr>
<td>4</td>
<td>Digital images/video, human vision, video quality</td>
<td>Chapter 2</td>
</tr>
<tr>
<td>5</td>
<td>Vector-matrix notation, image models, formulation of ill-posed problems in image/video processing</td>
<td>Appendix A, Appendix B</td>
</tr>
<tr>
<td>6</td>
<td>Decimation, interpolation, multi-resolution pyramids</td>
<td>3.2</td>
</tr>
<tr>
<td>7</td>
<td>Gradient estimation, edge-corner detection</td>
<td>3.3</td>
</tr>
<tr>
<td>8</td>
<td>Image enhancement, point operations, unsharp masking, bilateral filtering</td>
<td>3.1, 3.4</td>
</tr>
<tr>
<td>9</td>
<td>Noise filtering: LSI filters; adaptive, nonlinear, and non-local filters</td>
<td>3.5</td>
</tr>
<tr>
<td>10</td>
<td>Image restoration: iterative methods, POCS</td>
<td>3.6</td>
</tr>
<tr>
<td>11</td>
<td>Motion modeling, optical flow, correspondence</td>
<td>4.1, 4.2, 4.3</td>
</tr>
<tr>
<td>12</td>
<td>Differential methods: Lukas–Kanade, parametric models</td>
<td>4.4</td>
</tr>
<tr>
<td>13</td>
<td>Block matching, feature matching for parametric model estimation, phase-correlation method</td>
<td>4.5, 4.7</td>
</tr>
<tr>
<td>14</td>
<td>3D motion estimation, epipolar geometry</td>
<td>4.8</td>
</tr>
<tr>
<td>15</td>
<td>Change detection, video segmentation</td>
<td>5.2, 5.3</td>
</tr>
<tr>
<td>16</td>
<td>Motion tracking</td>
<td>5.4, 5.5</td>
</tr>
<tr>
<td>17</td>
<td>Motion-compensated filtering, multi-frame de-interlacing, de-noising</td>
<td>6.1, 6.2, 6.3</td>
</tr>
<tr>
<td>18</td>
<td>Super-resolution</td>
<td>6.5</td>
</tr>
<tr>
<td>19</td>
<td>Introduction to data/image compression, information theoretic concepts, entropy coding, arithmetic coding</td>
<td>7.1</td>
</tr>
<tr>
<td>20</td>
<td>Lossless bitplane coding, group 3/4, JBIG standards</td>
<td>7.2</td>
</tr>
<tr>
<td>21</td>
<td>Predictive data coding, JPEG-LS standard</td>
<td>7.2</td>
</tr>
<tr>
<td>22</td>
<td>DCT and JPEG image compression</td>
<td>7.3</td>
</tr>
<tr>
<td>23</td>
<td>Wavelet transform, JPEG-2000 image compression</td>
<td>7.4</td>
</tr>
<tr>
<td>24</td>
<td>MC-DCT, MPEG-1, MPEG-2</td>
<td>8.1, 8.2</td>
</tr>
<tr>
<td>25</td>
<td>MPEG-4 AVC/H.264 standard</td>
<td>8.3</td>
</tr>
<tr>
<td>26</td>
<td>HEVC</td>
<td>8.4</td>
</tr>
<tr>
<td>27</td>
<td>Scalable video coding (SVC), DASH adaptive streaming, error-resilience</td>
<td>8.5</td>
</tr>
<tr>
<td>28</td>
<td>3D/stereo and multi-view video compression</td>
<td>8.6</td>
</tr>
</tbody>
</table>
in various ISO and ITU groups over the years, where it is difficult to give individual credit to everyone.

Finally, I would like to express my gratitude to Xin Li (WVU), Eli Saber, Moncef Gabbouj, Janusz Konrad, and H. Joel Trussell for reviewing the manuscript at various stages. I would also like to thank Bernard Goodwin, Kim Boedigheimer, and Julie Nahil from Prentice Hall for their help and support.

—A. Murat Tekalp
Koç University
Istanbul, Turkey
April 2015
This page intentionally left blank
A. Murat Tekalp received a Ph.D. in electrical, computer, and systems engineering from Rensselaer Polytechnic Institute (RPI), Troy, New York, in 1984. He was with Eastman Kodak Company, Rochester, New York, from 1984 to 1987, and with the University of Rochester, Rochester, New York, from 1987 to 2005, where he was promoted to Distinguished University Professor. He is currently a professor at Koç University, Istanbul, Turkey. He served as the Dean of Engineering at Koç University from 2010 through 2013. His research interests are in the area of digital image and video processing, image and video compression, and video networking.

Dr. Tekalp is a fellow of IEEE and a member of Academia Europaea and Turkish Academy of Sciences. He received the TUBITAK Science Award (the highest scientific award in Turkey) in 2004. He is a former chair of the IEEE Technical Committee on Image and Multidimensional Signal Processing, and a founding member of the IEEE Technical Committee on Multimedia Signal Processing. He was appointed as the technical program co-chair for IEEE ICASSP 2000 in Istanbul, Turkey; the general chair of IEEE International Conference on Image Processing (ICIP) at Rochester, New York, in 2002; and technical program co-chair of EUSIPCO 2005 in Antalya, Turkey.

He was the editor-in-chief of the EURASIP journal Signal Processing: Image Communication (published by Elsevier) from 2000 through 2010. He also served as an associate editor for the IEEE Transactions on Signal Processing and IEEE Transactions on Image Processing. He was on the editorial board of IEEE’s Signal Processing Magazine (2007–2010). He is currently on the editorial board of the Proceedings of the IEEE. He also serves as a member of the European Research Council (ERC) Advanced Grant panels.
This page intentionally left blank
This page intentionally left blank
Digital images and video refer to 2D or 3D still and moving (time-varying) visual information, respectively. A still image is a 2D/3D spatial distribution of intensity that is constant with respect to time. A video is a 3D/4D spatio-temporal intensity pattern, i.e., a spatial-intensity pattern that varies with time. Another term commonly used for video is image sequence, since a video is represented by a time sequence of still images (pictures). The spatio-temporal intensity pattern of this time sequence of images is ordered into a 1D analog or digital video signal as a function of time only according to a progressive or interlaced scanning convention.

We begin with a short introduction to human visual perception and color models in Section 2.1. We give a brief review of analog-video representations in Section 2.2, mainly to provide a historical perspective. Next, we present 2D digital video representations and a brief summary of current standards in Section 2.3. We introduce 3D digital video display, representations, and standards in Section 2.4. Section 2.5 provides an overview of popular digital video applications, including digital TV, digital cinema, and video streaming. Finally, Section 2.6 discusses factors affecting video quality and quantitative and subjective video-quality assessment.
Chapter 2. Digital Images and Video

2.1 Human Visual System and Color

Video is mainly consumed by the human eye. Hence, many imaging system design choices and parameters, including spatial and temporal resolution as well as color representation, have been inspired by or selected to imitate the properties of human vision. Furthermore, digital image/video-processing operations, including filtering and compression, are generally designed and optimized according to the specifications of the human eye. In most cases, details that cannot be perceived by the human eye are regarded as irrelevant and referred to as perceptual redundancy.

2.1.1 Color Vision and Models

The human eye is sensitive to the range of wavelengths between 380 nm (blue end of the visible spectrum) and 780 nm (red end of the visible spectrum). The cornea, iris, and lens comprise an optical system that forms images on the retinal surface. There are about 100-120 million rods and 7-8 million cones in the retina [Wan 95, Fer 01]. They are receptor nerve cells that emit electrical signals when light hits them. The region of the retina with the highest density of photoreceptors is called the fovea. Rods are sensitive to low-light (scotopic) levels but only sense the intensity of the light; they enable night vision. Cones enable color perception and are best in bright (photopic) light. They have bandpass spectral response. There are three types of cones that are more sensitive to short (S), medium (M), and long (L) wavelengths, respectively. The spectral response of S-cones peak at 420 nm, M-cones at 534 nm, and L-cones at 564 nm, with significant overlap in their spectral response ranges and varying degrees of sensitivity at these range of wavelengths specified by the function $m_k(\lambda), k = r, g, b$, as depicted in Figure 2.1(a).

The perceived color of light $f(x_1, x_2, \lambda)$ at spatial location $(x_1, x_2)$ depends on the distribution of energy in the wavelength $\lambda$ dimension. Hence, color sensation can be achieved by sampling $\lambda$ into three levels to emulate color sensation of each type of cones as:

$$f_k(x_1, x_2) = \int f(x_1, x_2, \lambda) m_k(\lambda) d\lambda \quad k = r, g, b$$

(2.1)

where $m_k(\lambda)$ is the wavelength sensitivity function (also known as the color-matching function) of the $k$th cone type or color sensor. This implies that perceived color at any location $(x_1, x_2)$ depends only on three values $f_r$, $f_g$, and $f_b$, which are called the tristimulus values.

It is also known that the human eye has a secondary processing stage whereby the R, G, and B values sensed by the cones are converted into a luminance and two
2.1 Human Visual System and Color

color-difference (chrominance) values [Fer 01]. The luminance $Y$ is related to the perceived brightness of the light and is given by

$$Y(x_1, x_2) = \int f(x_1, x_2, \lambda) I(\lambda) d\lambda$$  \hspace{1cm} (2.2)

where $I(\lambda)$ is the International Commission on Illumination (CIE) luminous efficiency function, depicted in Figure 2.1(b), which shows the contribution of energy at each wavelength to a standard human observer’s perception of brightness. Two chrominance values describe the perceived color of the light. Color representations for color image processing are further discussed in Section 2.3.3.

![Figure 2.1](image)

**Figure 2.1** Spectral sensitivity: (a) CIE 1931 color-matching functions for a standard observer with a 2-degree field of view, where the curves $x$, $y$, and $z$ may represent $m_r(\lambda)$, $m_g(\lambda)$, and $m_b(\lambda)$, respectively, and (b) the CIE luminous efficiency function $I(\lambda)$ as a function of wavelength $\lambda$. 
Now that we have established that the human eye perceives color in terms of three component values, the next question is whether all colors can be reproduced by mixing three primary colors. The answer to this question is yes in the sense that most colors can be realized by mixing three properly chosen primary colors. Hence, inspired by human color perception, digital representation of color is based on the tri-stimulus theory, which states that all colors can be approximated by mixing three additive primaries, which are described by their color-matching functions. As a result, colors are represented by triplets of numbers, which describe the weights used in mixing the three primaries. All colors that can be reproduced by a combination of three primary colors define the color gamut of a specific device. There are different choices for selecting primaries based on additive and subtractive color models. We discuss the additive RGB and subtractive CMYK color spaces and color management in the following. However, an in-depth discussion of color science is beyond the scope of this book, and interested readers are referred to [Tru 93, Sha 98, Dub 10].

**RGB and CMYK Color Spaces**

The RGB model, inspired by human vision, is an additive color model in which red, green, and blue light are added together to reproduce a variety of colors. The RGB model applies to devices that capture and emit color light such as digital cameras, video projectors, LCD/LED TV and computer monitors, and mobile phone displays. Alternatively, devices that produce materials that reflect light, such as color printers, are governed by the subtractive CMYK (Cyan, Magenta, Yellow, Black) color model. Additive and subtractive color spaces are depicted in Figure 2.2. RGB and CMYK are *device-dependent* color models: i.e., different devices detect or reproduce a given RGB value differently, since the response of color elements (such as filters or dyes) to individual R, G, and B levels may vary among different manufacturers. Therefore, the RGB color model itself does not define absolute *red*, *green*, and *blue* (hence, the result of mixing them) colorimetrically.

When the exact chromaticities of red, green, and blue primaries are defined, we have a *color space*. There are several color spaces, such as CIE RGB, CIE XYZ, or sRGB. CIE RGB and CIE XYZ are the first formal color spaces defined by the CIE in 1931. Since display devices can only generate non-negative primaries, and an adequate amount of luminance is required, there is, in practice, a limitation on the gamut of colors that can be reproduced on a given device. Color characteristics of a device can be specified by its International Color Consortium (ICC) profile.
Color Management

Color management must be employed to generate the exact same color on different devices, where the device-dependent color values of the input device, given its ICC profile, is first mapped to a standard device-independent color space, sometimes called the Profile Connection Space (PCS), such as CIEXYZ. They are then mapped to the device-dependent color values of the output device given the ICC profile of the output device. Hence, an ICC profile is essentially a mapping from a device color space to the PCS and from the PCS to a device color space. Suppose we have particular RGB and CMYK devices and want to convert the RGB values to CMYK. The first step is to obtain the ICC profiles of concerned devices. To perform the conversion, each (R, G, B) triplet is first converted to the PCS using the ICC profile of the RGB device. Then, the PCS is converted to the C, M, Y, and K values using the profile of the second device.

Color management may be side-stepped by calibrating all devices to a common standard color space, such as sRGB, which was developed by HP and Microsoft in 1996. sRGB uses the color primaries defined by the ITU-R recommendation BT.709, which standardizes the format of high-definition television. When such a calibration is done well, no color translations are needed to get all devices to handle colors consistently. Avoiding the complexity of color management was one of the goals in developing sRGB [IEC 00].

2.1.2 Contrast Sensitivity

Contrast can be defined as the difference between the luminance of a region and its background. The human visual system is more sensitive to contrast than absolute
luminance; hence, we can perceive the world around us similarly regardless of changes in illumination. Since most images are viewed by humans, it is important to understand how the human visual system senses contrast so that algorithms can be designed to preserve the more visible information and discard the less visible ones. Contrast-sensitivity mechanisms of human vision also determine which compression or processing artifacts we see and which we don’t. The ability of the eye to discriminate between changes in intensity at a given intensity level is quantified by Weber’s law.

**Weber’s Law**

Weber’s law states that smaller intensity differences are more visible on a darker background and can be quantified as

\[
\frac{\Delta I}{I} = c \text{ (constant), for } I > 0
\]

where \(\Delta I\) is the just noticeable difference (JND) [Gon 07]. Eqn. (2.5) states that the JND grows proportional to the intensity level \(I\). Note that \(I = 0\) denotes the darkest intensity, while \(I = 255\) is the brightest. The value of \(c\) is empirically found to be around 0.02. The experimental set-up to measure the JND is shown in Figure 2.3(a). The rods and cones comply with Weber’s law above \(-2.6\) log candelas (cd)/m² (moonlight) and \(2\) log cd/m² (indoor) luminance levels, respectively [Fer 01].

**Brightness Adaptation**

The human eye can adapt to different illumination/intensity levels [Fer 01]. It has been observed that when the background-intensity level the observer has adapted to is different from \(I\), the observer’s intensity resolution ability decreases. That is, when \(I_0\) is different from \(I\), as shown in Figure 2.3(b), the JND \(\Delta I\) increases relative to the case \(I_0 = I\). Furthermore, the simultaneous contrast effect illustrates that humans perceive the brightness of a square with constant intensity differently as the intensity of the background varies from light to dark [Gon 07].

It is also well-known that the human visual system undershoots and overshoots around the boundary of step transitions in intensity as demonstrated by the Mach band effect [Gon 07].

**Visual Masking**

Visual masking refers to a nonlinear phenomenon experimentally observed in the human visual system when two or more visual stimuli that are closely coupled in space or time are presented to a viewer. The action of one visual stimulus on the visibility of another is called masking. The effect of masking may be a decrease in
2.1 Human Visual System and Color

Spatial Masking
Spatial masking is observed when a viewer is presented with a superposition of a target pattern and mask (background) image [Fer 01]. The effect states that the visibility of the target pattern is lower when the background is spatially busy. Spatial busyness measures include local image variance or textureness. Spatial masking implies that visibility of noise or artifact patterns is lower in spatially busy areas of an image as compared to spatially uniform image areas.

Temporal Masking
Temporal masking is observed when two stimuli are presented sequentially [Bre 07]. Salient local changes in luminance, hue, shape, or size may become undetectable in the presence of large coherent object motion [Suc 11]. Considering video frames as a sequence of stimuli, fast-moving objects and scene cuts can trigger a temporal-masking effect.

2.1.3 Spatio-Temporal Frequency Response
An understanding of the response of the human visual system to spatial and temporal frequencies is important to determine video-system design parameters and video-compression parameters, since frequencies that are invisible to the human eye are irrelevant.
Spatial-Frequency Response

Spatial frequencies are related to how still (static) image patterns vary in the horizontal and vertical directions in the spatial plane. The spatial-frequency response of the human eye varies with the viewing distance; i.e., the closer we get to the screen the better we can see details. In order to specify the spatial frequency independent of the viewing distance, spatial frequency (in cycles/distance) must be normalized by the viewing distance \( d \), which can be done by defining the viewing angle \( \theta \) as shown in Figure 2.4(a).

Let \( w \) denote the picture width. If \( w/2 \ll d \), then \( \frac{\theta}{2} \approx \sin \frac{\theta}{2} = \frac{w/2}{d} \), considering the right triangle formed by the viewer location, an end of the picture, and the middle of the picture. Hence,

\[
\theta \approx \frac{w}{d} \text{ (radians)} = \frac{180w}{\pi d} \text{ (degrees)} \quad (2.3)
\]

Let \( f_w \) denote the number of cycles per picture width, then the normalized horizontal spatial frequency (i.e., number of cycles per viewing degree) \( f_\theta \) is given by

\[
f_\theta = \frac{f_w}{\theta} = \frac{f_w d}{w} \text{ (cycles / radian)} = \frac{\pi d f_w}{180 w} \text{ (cycles / degree)} \quad (2.4)
\]

The normalized vertical spatial frequency can be defined similarly in the units of cycles/degree. As we move away from the screen \( d \) increases, and the same number of cycles per picture width \( f_w \) appears as a larger frequency \( f_\theta \) per viewing degree. Since the human eye has reduced contrast sensitivity at higher frequencies, the same pattern is more difficult to see from a larger distance \( d \). The horizontal and vertical resolution (number of pixels and lines) of a TV has been determined such that horizontal and vertical sampling frequencies are twice the highest frequency we can see (according to the Nyquist sampling theorem), assuming a fixed value for the ratio \( d/w \)—i.e., viewing distance over picture width. Given a fixed viewing distance, clearly we need more video resolution (pixels and lines) as picture (screen) size increases to experience the same video quality.

Figure 2.4(b) shows the spatial-frequency response, which varies by the average luminance level, of the eye for both the luminance and chrominance components of still images. We see that the spatial-frequency response of the eye, in general, has low-pass/band-pass characteristics, and our eyes are more sensitive to higher frequency patterns in the luminance components compared with those in the chrominance components. The latter observation is the basis of the conversion from RGB to the luminance-chrominance space for color image processing and the reason we subsample the two chrominance components in color image/video compression.
Temporal-Frequency Response

Video is displayed as a sequence of still frames. The frame rate is measured in terms of the number of pictures (frames) displayed per second or Hertz (Hz). The frame rates for cinema, television, and computer monitors have been determined according to the temporal-frequency response of our eyes. The human eye has lower sensitivity to higher temporal frequencies due to temporal integration of incoming light into the retina, which is also known as vision persistence. It is well known that the integration period is inversely proportional to the incoming light intensity. Therefore, we can see higher temporal frequencies on brighter screens. Psycho-visual experiments indicate the human eye cannot perceive flicker if the refresh rate of the display (temporal frequency) is more than 50 times per second for TV screens. Therefore, the frame rate for TV is set at 50-60 Hz, while the frame rate for brighter computer monitors is 72 Hz or higher, since the brighter the screen the higher the critical flicker frequency.

Interaction Between Spatial- and Temporal-Frequency Response

Video exhibits both spatial and temporal variations, and spatial- and temporal-frequency responses of the eye are not mutually independent. Hence, we need to understand the spatio-temporal frequency response of the eye. The effects of changing average luminance on the contrast sensitivity for different combinations of spatial and temporal frequencies have been investigated [Nes 67]. Psycho-visual experiments
indicate that when the temporal (spatial) frequencies are close to zero, the spatial (temporal) frequency response has bandpass characteristics. At high temporal (spatial) frequencies, the spatial (temporal) frequency response has low-pass characteristics with smaller cut-off frequency as temporal (spatial) frequency increases. This implies that we can exchange spatial video resolution for temporal resolution, and vice versa. Hence, when a video has high motion (moves fast), the eyes cannot sense high spatial frequencies (details) well if we exclude the effect of eye movements.

Eye Movements
The human eye is similar to a sphere that is free to move like a ball in a socket. If we look at a nearby object, the two eyes turn in; if we look to the left, the right eye turns in and the left eye turns out; if we look up or down, both eyes turn up or down together. These movements are directed by the brain [Hub 88]. There are two main types of gaze-shifting eye movements, saccadic and smooth pursuit, that affect the spatial- and spatio-temporal frequency response of the eye. Saccades are rapid movements of the eyes while scanning a visual scene. “Saccadic eye movements” enable us to scan a greater area of the visual scene with the high-resolution fovea of the eye. On the other hand, “smooth pursuit” refers to movements of the eye while tracking a moving object, so that a moving image remains nearly static on the high-resolution fovea. Obviously, smooth pursuit eye movements affect the spatio-temporal frequency response of the eye. This effect can be modeled by tracking eye movements of the viewer and motion compensating the contrast sensitivity function accordingly.

2.1.4 Stereo/Depth Perception
Stereoscopy creates the illusion of 3D depth from two 2D images, a left and a right image that we should view with our left and right eyes. The horizontal distance between the eyes (called \textit{interpupilar distance}) of an average human is 6.5 cm. The difference between the left and right retinal images is called \textit{binocular disparity}. Our brain deducts depth information from this binocular disparity. 3D display technologies that enable viewing of right and left images with our right and left eyes, respectively, are discussed in Section 2.4.1.

Accommodation, Vergence, and Visual Discomfort
In human stereo vision, there are two oculomotor mechanisms, accommodation (where we focus) and vergence (where we look), which are reflex eye movements. Accommodation is the process by which the eye changes optical focus to maintain a clear image of an object as its distance from the eye varies. Vergence or convergence
are the movements of both eyes to make sure the image of the object being looked at falls on the corresponding spot on both retinas. In real 3D vision, accommodation and vergence distances are the same. However, in flat 3D displays both left and right images are displayed on the plane of the screen, which determines the accommodation distance, while we look and perceive 3D objects at a different distance (usually closer to us), which is the vergence distance. This difference between accommodation and vergence distances may cause serious discomfort if it is greater than some tolerable amount. The depth of an object in the scene is determined by the disparity value, which is the displacement of a feature point between the right and left views. The depth, hence the difference between accommodation and vergence distances, can be controlled by 3D-video (disparity) processing at the content preparation stage to provide a comfortable 3D viewing experience.

Another cause of viewing discomfort is the cross-talk between the left and right views, which may cause ghosting and blurring. Cross-talk may result from imperfections in polarizing filters (passive glasses) or synchronization errors (active shutters), but it is more prominent in auto-stereoscopic displays where the optics may not completely prevent cross-talk between the left and right views.

### Binocular Rivalry/Suppression Theory

Binocular rivalry is a visual perception phenomenon that is observed when different images are presented to right and left eyes [Wad 96]. When the quality difference between the right and left views are small, according to the suppression theory of stereo vision, the human eye can tolerate absence of high-frequency content in one of the views; therefore, two views can be represented at unequal spatial resolutions or quality. This effect has lead to asymmetric stereo-video coding, where only the dominant view is encoded with high fidelity (bitrate). The results have shown that perceived 3D-video quality of such asymmetric processed stereo pairs is similar to that of symmetrically encoded sequences at higher total bitrate. They also observe that scaling (zoom in/out) one or both views of a stereoscopic test sequence does not affect depth perception. We note that these results have been confirmed on short test sequences. It is not known whether asymmetric view resolution or quality would cause viewing discomfort over longer videos with increased period of viewing.

### 2.2 Analog Video

We used to live in a world of analog images and video, where we dealt with photographic film, analog TV sets, videocassette recorders (VCRs), and camcorders.
Chapter 2. Digital Images and Video

For video distribution, we relied on analog TV broadcasts and analog cable TV, which transmitted predetermined programming at a fixed rate. Analog video, due to its nature, provided a very limited amount of interactivity, e.g., only channel selection on the TV and fast-forward search and slow-motion replay on the VCR. Additionally, we had to live with the NTSC/PAL/SECAM analog signal formats with their well-known artifacts and very low still-frame image quality. In order to display NTSC signals on computer monitors or European TV sets, we needed expensive transcoders. In order to display a smaller version of the NTSC picture in a corner of the monitor, we first had to digitize the whole picture and then digitally reduce its size. Searching a video archive for particular footage required tedious visual scanning of a whole bunch of videotapes. Motion pictures were recorded on photographic film, which is a high-resolution analog medium, or on laser discs as analog signals using optical technology. Manipulation of analog video is not an easy task, since it requires digitization of the analog signal into digital form first.

Today almost all video capture, processing, transmission, storage, and search are in digital form. In this section, we describe the nature of the analog-video signal because an understanding of history of video and the limitations of analog video formats is important. For example, interlaced scanning originates from the history of analog video. We note that video digitized from analog sources is limited by the resolution and the artifacts of the respective analog signal.

2.2.1 Progressive vs. Interlaced Scanning

The analog-video signal refers to a one-dimensional (1D) signal \( s(t) \) of time that is obtained by sampling \( s(x_1, x_2, t) \) in the vertical \( x_2 \) and temporal coordinates. This conversion of 3D spatio-temporal signal into a 1D temporal signal by periodic vertical-temporal sampling is called scanning. The signal \( s(t) \), then, captures the time-varying image intensity \( s(x_1, x_2, t) \) only along the scan lines. It also contains the timing information and blanking signals needed to align pictures.

The most commonly used scanning methods are progressive scanning and interlaced scanning. Progressive scan traces a complete picture, called a frame, at every \( \Delta t \) sec. The spot flies back from B to C, called the horizontal retrace, and from D to A, called the vertical retrace, as shown in Figure 2.5(a). For example, the computer industry uses progressive scanning with \( \Delta t = 1/72 \) sec for monitors. On the other hand, the TV industry uses 2:1 interlaced scan where the odd-numbered and even-numbered lines, called the odd field and the even field, respectively, are traced in turn. A 2:1 interlaced scanning raster is shown in Figure 2.5(b), where the solid line
and the dotted line represent the odd and the even fields, respectively. The spot snaps back from D to E, and from F to A, for even and odd fields, respectively, during the vertical retrace intervals.

2.2.2 Analog-Video Signal Formats

Some important parameters of the video signal are the vertical resolution, aspect ratio, and frame/field rate. The vertical resolution is related to the number of scan lines per frame. The aspect ratio is the ratio of the width to the height of a frame. As discussed in Section 2.1.3, the human eye does not perceive flicker if the refresh rate of the display is more than 50 Hz. However, for analog TV systems, such a high frame rate, while preserving the vertical resolution, requires a large transmission bandwidth. Thus, it was determined that analog TV systems should use interlaced scanning, which trades vertical resolution to reduced flickering within a fixed bandwidth.

An example analog-video signal $s(t)$ is shown in Figure 2.6. Blanking pulses (black) are inserted during the retrace intervals to blank out retrace lines on the monitor. Sync pulses are added on top of the blanking pulses to synchronize the receiver’s horizontal and vertical sweep circuits. The sync pulses ensure that the picture starts at the top-left corner of the receiving monitor. The timing of the sync pulses is, of course, different for progressive and interlaced video.

Several analog-video signal standards, which are obsolete today, have different image parameters (e.g., spatial and temporal resolution) and differ in the way they handle color. These can be grouped as: i) component analog video; ii) composite video; and iii) S-video (Y/C video). Component analog video refers to individual
red (R), green (G), and blue (B) video signals. Composite-video format encodes the chrominance components on top of the luminance signal for distribution as a single signal that has the same bandwidth as the luminance signal. Different composite-video formats, e.g., NTSC (National Television Systems Committee), PAL (Phase Alternation Line), and SECAM (Systeme Electronique Color Avec Memoire), have been used in different regions of the world. The composite signal usually results in errors in color rendition, known as hue and saturation errors, because of inaccuracies in the separation of the color signals. S-video is a compromise between the composite video and component video, where we represent the video with two component signals, a luminance and a composite chrominance signal. The chrominance signals have been based on (I,Q) or (U,V) representation for NTSC, PAL, or SECAM systems. S-video was used in consumer-quality videocassette recorders and analog camcorders to obtain image quality better than that of composite video. Cameras specifically designed for analog television pickup from motion picture film were called telecine cameras. They employed frame-rate conversion from 24 frames/sec to 60 fields/sec.

2.2.3 Analog-to-Digital Conversion

The analog-to-digital (A/D) conversion process consists of pre-filtering (for anti-aliasing), sampling, and quantization of component (R, G, B) signal or composite signal. The ITU (International Telecommunications Union) and SMPTE (Society of Motion Picture and Television Engineers) have standardized sampling parameters for both component and composite video to enable easy exchange of digital video
across different platforms. For A/D conversion of component signals, the horizontal sampling rate of 13.5 MHz for the luma component and 6.75 MHz for two chroma components were chosen, because they satisfy the following requirements:

1. Minimum sampling frequency (Nyquist rate) should be $4.2 \times 2 = 8.4$ MHz for 525/30 NTSC luma and $5 \times 2 = 10$ MHz for 625/50 PAL luma signals.
2. Sampling rate should be an integral multiple of the line rate, so samples in successive lines are correctly aligned (on top of each other).
3. For sampling component signals, there should be a single rate for 525/30 and 625/50 systems; i.e., the sampling rate should be an integral multiple of line rates (lines/sec) of both $29.97 \times 525 = 15,734$ and $25 \times 625 = 15,625$.

For sampling the composite signal, the sampling frequency must be an integral multiple of the sub-carrier frequency to simplify composite signal to RGB decoding of sampled signal. It is possible to operate at 3 or 4 times the subcarrier frequency, although most systems choose to employ $4 \times 3.58 = 14.32$ MHz for NTSC and $4 \times 4.43 = 17.72$ MHz for PAL signals, respectively.

### 2.3 Digital Video

We have experienced a digital media revolution in the last couple of decades. TV and cinema have gone all-digital and high-definition, and most movies and some TV broadcasts are now in 3D format. High-definition digital video has landed on laptops, tablets, and cellular phones with high-quality media streaming over the Internet. Apart from the more robust form of the digital signal, the main advantage of digital representation and transmission is that they make it easier to provide a diverse range of services over the same network. Digital video brings broadcasting, cinema, computers, and communications industries together in a truly revolutionary manner, where telephone, cable TV, and Internet service providers have become fierce competitors. A single device can serve as a personal computer, a high-definition TV, and a videophone. We can now capture live video on a mobile device, apply digital processing on a laptop or tablet, and/or print still frames at a local printer. Other applications of digital video include medical imaging, surveillance for military and law enforcement, and intelligent highway systems.

#### 2.3.1 Spatial Resolution and Frame Rate

Digital-video systems use component color representation. Digital color cameras provide individual RGB component outputs. Component color video avoids the
artifacts that result from analog composite encoding. In digital video, there is no need for blanking or sync pulses, since it is clear where a new line starts given the number of pixels per line.

The horizontal and vertical resolution of digital video is related to the pixel sampling density, i.e., the number of pixels per unit distance. The number of pixels per line and the number of lines per frame is used to classify video as standard, high, or ultra-high definition, as depicted in Figure 2.7. In low-resolution digital video, pixellation (aliasing) artifact arises due to lack of sufficient spatial resolution. It manifests itself as jagged edges resulting from individual pixels becoming visible. The visibility of pixellation artifacts varies with the size of the display and the viewing distance. This is quite different from analog video where the lack of spatial-resolution results in blurring of image in the respective direction.

The frame/field rate is typically 50/60 Hz, although some displays use frame interpolation to display at 100/120, 200 or even 400 Hz. The notation 50i (or 60i) indicates interlaced video with 50 (60) fields/sec, which corresponds to 25 (30) pictures/sec obtained by weaving the two fields together. On the other hand, 50p (60p) denotes 50 (60) full progressive frames/sec.

The arrangement of pixels and lines in a contiguous region of the memory is called a bitmap. There are five key parameters of a bitmap: the starting address in the memory, the number of pixels per line, the pitch value, the number of lines, and the number of bits per pixel. The pitch value specifies the distance in memory from the start of one line to the next. The most common use of pitch different from

![Figure 2.7 Digital-video spatial-resolution formats.](image-url)
the number of pixels per line is to set pitch to the next highest power of 2, which may help certain applications run faster. Also, when dealing with interlaced inputs, setting the pitch to double the number of pixels per line facilitates writing lines from each field alternately in memory. This will form a “weaved frame” in a contiguous region of the memory.

2.3.2 Color, Dynamic Range, and Bit-Depth
This section addresses color representation, dynamic range, and bit-depth in digital images/video.

Color Capture and Display
Color cameras can be the three-sensor type or single-sensor type. Three-sensor cameras capture R, G, and B components using different CCD panels, using an optical beam splitter; however, they may suffer from synchronicity problems and high cost, while single-sensor cameras often have to compromise spatial resolution. This is because a color filter array is used so that each CCD element captures one of R, G, or B pixels in some periodic pattern. A commonly used color filter pattern is the Bayer array, shown in Figure 2.8, where two out of every four pixels are green, one is red, and one is blue, since green signal contributes the most to the luminance channel. The missing pixel values in each color channel are computed by linear or adaptive

![Bayer color-filter array pattern.](image-url)
interpolation filters, which may result in some aliasing artifacts. Similar color filter array patterns are also employed in LCD/LED displays, where the human eye performs low-pass filtering to perceive a full-colored image.

**Dynamic Range**

The dynamic range of a capture device (e.g., a camera or scanner) or a display device is the ratio between the maximum and minimum light intensities that can be represented. The luminance levels in the environment range from \(-4 \log \text{cd/m}^2\) (starlight) to \(6 \log \text{cd/m}^2\) (sunlight); i.e., the dynamic range is about 10 log units [Fer 01]. The human eye has complex fast and slow adaptation schemes to cope with this large dynamic range. However, a typical imaging device (camera or display) has a maximum dynamic range of 300:1, which corresponds to 2.5 log units. Hence, our ability to capture and display a foreground object subject to strong backlighting with proper contrast is limited. High dynamic range (HDR) imaging aims to remedy this problem.

**HDR Image Capture**

HDR image capture with a standard dynamic range camera requires taking a sequence of pictures at different exposure levels, where raw pixel exposure data (linear in exposure time) are combined by weighted averaging to obtain a single HDR image [Gra 10]. There are two possible ways to display HDR images: i) employ new higher dynamic range display technologies, or ii) employ local tone-mapping algorithms for dynamic range compression (see Chapter 3) to better render details in bright or dark areas on a standard display [Rei 07].

**HDR Displays**

Recently, new display technologies that are capable of up to 50,000:1 or 4.7 log units dynamic range with maximum intensity 8500 cd/m², compared to standard displays with contrast ratio 2 log units and maximum intensity 300 cd/m², have been proposed [See 04]. This high dynamic range matches the human eye’s short time-scale (fast) adaptation capability well, which enables our eyes to capture approximately 5 log units of dynamic range at the same time.

**Bit-Depth**

Image-intensity values at each sample are quantized for a finite-precision representation. Today, each color component signal is typically represented with 8 bits per pixel, which can capture 255:1 dynamic range for a total of 24 bits/pixel and \(2^{24}\)
2.3 Digital Video

distinct colors to avoid “contouring artifacts.” Contouring results in slowly varying regions of image intensity due to insufficient bit resolution. Some applications, such as medical imaging and post-production editing of motion pictures may require 10, 12, or more bits/pixel/color. In high dynamic range imaging, 16 bits/pixel/color is required to capture a 50,000:1 dynamic range, which is now supported in JPEG.

Digital video requires much higher data rates and transmission bandwidths as compared to digital audio. CD-quality digital audio is represented with 16 bits/sample, and the required sampling rate is 44 kHz. Thus, the resulting data rate is approximately 700 kbits/sec (kbps). This is multiplied by 2 for stereo audio. In comparison, a high-definition TV signal has 1920 pixels/line and 1080 lines for each luminance frame, and 960 pixels/line and 540 lines for each chrominance frame. Since we have 25 frames/sec and 8 bits/pixel/color, the resulting data rate exceeds 700 Mbps, which testifies to the statement that a picture is worth 1000 words! Thus, the feasibility of digital video is dependent on image-compression technology.

2.3.3 Color Image Processing

Color images/video are captured and displayed in the RGB format. However, they are often converted to an intermediate representation for efficient compression and processing. We review the luminance-chrominance (for compression and filtering) and the normalized RGB and hue-saturation-intensity (HSI) (for color-specific processing) representations in the following.

Luminance-Chrominance

The luminance-chrominance color model was used to develop an analog color TV transmission system that is backwards compatible with the legacy analog black and white TV systems. The luminance component, denoted by Y, corresponds to the gray-level representation of video, while the two chrominance components, denoted by U and V for analog video or Cr and Cb for digital video, represent the deviation of color from the gray level on blue–yellow and red–cyan axes. It has been observed that the human visual system is less sensitive to variations (higher frequencies) in chrominance components (see Figure 2.4(b)). This has resulted in the subsampled chrominance formats, such as 4:2:2 and 4:2:0. In the 4:2:2 format, the chrominance components are subsampled only in the horizontal direction, while in 4:2:0 they are subsampled in both directions as illustrated in Figure 2.9. The luminance-chrominance representation offers higher compression efficiency, compared to the RGB representation due to this subsampling.
ITU-R BT.709 defines the conversion between RGB and YCrCb representations as:

\[
\begin{align*}
Y &= 0.299 R + 0.587 G + 0.114 B \\
Cr &= 0.499 R - 0.418 G - 0.0813 B + 128 \\
Cb &= -0.169 R - 0.331 G + 0.499 B + 128
\end{align*}
\] (2.6)

which states that the human visual system perceives the contribution of R-G-B to image intensity approximately with a 3-6-1 ratio, i.e., red is weighted by 0.3, green by 0.6 and blue by 0.1.

The inverse conversion is given by

\[
\begin{align*}
R &= Y + 1.402 (Cr - 128) \\
G &= Y - 0.714 (Cr - 128) - 0.344 (Cb - 128) \\
B &= Y + 1.772 (Cb - 128)
\end{align*}
\] (2.7)

The resulting R, G, and B values must be truncated to the range (0, 255) if they fall outside. We note that Y-Cr-Cb is not a color space. It is a way of encoding the RGB information, and actual colors displayed depends on the specific RGB space used.

A common practice in color image processing, such as edge detection, enhancement, denoising, restoration, etc., in the luminance-chrominance domain is to process only the luminance (Y) component of the image. There are two main reasons for this: i) processing R, G, and B components independently may alter the color balance of the image, and ii) the human visual system is not very sensitive to high frequencies in the chrominance components. Therefore, we first convert a color image

![Chrominance subsampling formats](image)
into Y-Cr-Cb color space, then perform image enhancement, denoising, restoration, etc., on the Y channel only. We then transform the processed Y channel and unprocessed Cr and Cb channels back to the R-G-B domain for display.

**Normalized rgb**

Normalized rgb components aim to reduce the dependency of color represented by the RGB values on image brightness. They are defined by

\[
\begin{align*}
    r &= R / (R + G + B) \\
    g &= G / (R + G + B) \\
    b &= B / (R + G + B)
\end{align*}
\]  

(2.8)

The normalized \(r, g, b\) values are always within the range 0 to 1, and

\[
r + g + b = 1
\]

(2.9)

Hence, they can be specified by any two components, typically by \((r, g)\) and the third component can be obtained from Eqn. (2.9). The normalized rgb domain is often used in color-based object detection, such as skin-color or face detection.

**Example.** We demonstrate how the normalized rgb domain helps to detect similar colors independent of brightness by means of an example: Let’s assume we have two pixels with (R, G, B) values (230, 180, 50) and (115, 90, 25). It is clear that the second pixel is half as bright as the first, which may be because it is in a shadow. In the normalized rgb, both pixels are represented by \(r = 0.50, g = 0.39,\) and \(b = 0.11\). Hence, it is apparent that they represent the same color after correcting for brightness difference by the normalization.

**Hue-Saturation-Intensity (HSI)**

Color features that best correlate with human perception of color are hue, saturation, and intensity. Hue relates to the dominant wavelength, saturation relates to the spread of power about this wavelength (purity of the color), and intensity relates to the perceived luminance (similar to the Y channel). There is a family of color spaces that specify colors in terms of hue, saturation, and intensity, known as HSI spaces. Conversion to HSI where each component is in the range \([0,1]\) can be performed from the scaled RGB, where each component is divided by 255 so they are in the
range [0,1]. The HSI space specifies color in cylindrical coordinates and conversion formulas (2.10) are nonlinear [Gon 07].

\[
H = \begin{cases} 
\theta & \text{if } B \leq G \\
360 - \theta & \text{if } B > G 
\end{cases}
\]

where \( \theta = \arccos\left\{ \frac{1/2[(R - G) + (R - B)]}{\sqrt{(R - G)^2 + (R - B)(G - B)}} \right\} \)

\[
S = 1 - \frac{3\min\{R,G,B\}}{R + G + B}
\]

\[
I = \frac{R + G + B}{3}
\]  

(2.10)

Note that HSI is not a perceptually uniform color space, i.e., equal perturbations in the component values do not result in perceptually equal color variations across the range of component values. The CIE has also standardized some perceptually uniform color spaces, such as L*, u*, v* and L*, a*, b* (CIELAB).

### 2.3.4 Digital-Video Standards

Exchange of digital video between different products, devices, and applications requires digital-video standards. We can group digital-video standards as video-format (resolution) standards, video-interface standards, and image/video compression standards. In the early days of analog TV, cinema (film), and cameras (cassette), the computer, TV, and consumer electronics industries established different display resolutions and scanning standards. Because digital video has brought cinema, TV, consumer electronics, and computer industries ever closer, standardization across industries has started. This section introduces recent standards and standardization efforts.

#### Video-Format Standards

Historically, standardization of digital-video formats originated from different sources: ITU-R driven by the TV industry, SMPTE driven by the motion picture industry, and computer/consumer electronics associations.

Digital video was in use in broadcast TV studios even in the days of analog TV, where editing and special effects were performed on digitized video because it is easier to manipulate digital images. Working with digital video avoids artifacts that would otherwise be caused by repeated analog recording of video on tapes during various production stages. Digitization of analog video has also been needed for conversion
between different analog standards, such as from PAL to NTSC, and vice versa. ITU-R (formerly CCIR) Recommendation BT.601 defines a standard definition TV (SDTV) digital-video format for 525-line and 625-line TV systems, also known as digital studio standard, which is originally intended to digitize analog TV signals to permit digital post-processing as well as international exchange of programs. This recommendation is based on component video with one luminance (Y) and two chrominance (Cr and Cb) signals. The sampling frequency for analog-to-digital (A/D) conversion is selected to be an integer multiple of the horizontal sweep frequencies (line rates) $f_{h,525} = 525 \times 29.97 = 15,734$ and $f_{h,625} = 625 \times 25 = 15,625$ in both 525- and 625-line systems, which was discussed in Section 2.2.3. Thus, for the luminance

$$f_{s,lum} = 858 \; f_{h,525} = 864 \; f_{h,625} = 13.5 \; MHz$$

i.e., 525 and 625 line systems have 858 and 864 samples/line, respectively, and for chrominance

$$f_{s,chr} = f_{s,lum}/2 = 6.75 \; MHz$$

ITU-R BT.601 standards for both 525- and 625-line SDTV systems employ interlaced scan, where the raw data rate is 165.9 Mbps. The parameters of both formats are shown in Table 2.1. Historically, interlaced SDTV was displayed on analog cathode ray tube (CRT) monitors, which employ interlaced scanning at 50/60 Hz. Today, flat-panel displays and projectors can display video at 100/120 Hz interlace or progressive mode, which requires scan-rate conversion and de-interlacing of the 50i/60i ITU-R BT.601 [ITU 11] broadcast signals.

Recognizing that the resolution of SDTV is well behind today’s technology, a new high-definition TV (HDTV) standard, ITU-R BT.709-5 [ITU 02], which doubles the resolution of SDTV in both horizontal and vertical directions, has been approved with three picture formats: 720p, 1080i, and 1080p. Table 2.1 shows their parameters. Today broadcasters use either 720p/50/60 (called HD) or 1080i/25/29.97 (called FullHD). There are no broadcasts in 1080p format at this time. Note that many 1080i/25 broadcasts use horizontal sub-sampling to 1440 pixels/line to save bitrate. 720p/50 format has full temporal resolution 50 progressive frames per second (with 720 lines). Note that most international HDTV events are captured in either 1080i/25 or 1080i/29.97 (for 60 Hz countries) and presenting 1080i/29.97
in 50 Hz countries or vice versa requires scan rate conversion. For 1080i/25 content, 720p/50 broadcasters will need to de-interlace the signal before transmission, and for 1080i/29.97 content, both de-interlacing and frame-rate conversion is required. Furthermore, newer 1920×1080 progressive scan consumer displays require upscaling 1280×720 pixel HD broadcast and 1440×1080i/25 sub-sampled FullHD broadcasts.

In the computer and consumer electronics industry, standards for video-display resolutions are set by a consortia of organizations such as Video Electronics Standards Association (VESA) and Consumer Electronics Association (CEA). The display standards can be grouped as Video Graphics Array (VGA) and its variants and Extended Graphics Array (XGA) and its variants. The favorite aspect ratio of the display industry has shifted from the earlier 4:3 to 16:10 and 16:9. Some of these standards are shown in Table 2.2. The refresh rate was an important parameter for CRT monitors. Since activated LCD pixels do not flash on/off between frames, LCD monitors do not exhibit refresh-induced flicker. The only part of an LCD monitor that can produce CRT-like flicker is its backlight, which typically operates at 200 Hz.

Recently, standardization across TV, consumer electronics, and computer industries has started, resulting in the so-called convergence enabled by digital video. For example, some laptops and cellular phones now feature 1920×1080 progressive mode, which is a format jointly supported by TV, consumer electronics, and computer industries.

Ultra-high definition television (UHDTV) is the most recent standard proposed by NHK Japan and approved as ITU-R BT.2020 [ITU 12]. It supports the 4K (2160p) and 8K (4320p) digital-video formats shown in Table 2.1. The Consumer Electronics Association announced that “ultra high-definition” or “ultra HD” or

<table>
<thead>
<tr>
<th>Table 2.1 ITU-R TV Broadcast Standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
</tr>
<tr>
<td>BT.601-7 480i</td>
</tr>
<tr>
<td>BT.601-7 576i</td>
</tr>
<tr>
<td>BT.709-5 720p</td>
</tr>
<tr>
<td>BT.709-5 1080i</td>
</tr>
<tr>
<td>BT.709-5 1080p</td>
</tr>
<tr>
<td>BT.2020 2160p</td>
</tr>
<tr>
<td>BT.2020 4320p</td>
</tr>
</tbody>
</table>
“UHD” would be used for displays that have an aspect ratio of at least 16:9 and at least one digital input capable of carrying and presenting native video at a minimum resolution of $3,840 \times 2,160$ pixels. The ultra-HD format is very similar to 4K digital cinema format (see Section 2.5.2) and may become an across industries standard in the near future.

**Video-Interface Standards**

Digital-video interface standards enable exchange of uncompressed video between various consumer electronics devices, including digital TV monitors, computer monitors, blu-ray devices, and video projectors over cable. Two such standards are Digital Visual Interface (DVI) and High-Definition Multimedia Interface (HDMI). HDMI is the most popular interface that enables transfer of video and audio on a single cable. It is backward compatible with DVI-D or DVI-I. HDMI 1.4 and higher support 2160p digital cinema and 3D stereo transfer.

**Image- and Video-Compression Standards**

Various digital-video applications, e.g., SDTV, HDTV, 3DTV, video on demand, interactive games, and videoconferencing, reach potential users over either broadcast channels or the Internet. Digital cinema content must be transmitted to movie theatres over satellite links or must be shipped in harddisks. Raw (uncompressed) data rates for digital video are prohibitive, since uncompressed broadcast HDTV requires

<table>
<thead>
<tr>
<th>Standard</th>
<th>Pixels</th>
<th>Lines</th>
<th>Aspect Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGA</td>
<td>640</td>
<td>480</td>
<td>4:3</td>
</tr>
<tr>
<td>WSVGA</td>
<td>1024</td>
<td>576</td>
<td>16:9</td>
</tr>
<tr>
<td>XGA</td>
<td>1024</td>
<td>768</td>
<td>4:3</td>
</tr>
<tr>
<td>WXGA</td>
<td>1366</td>
<td>768</td>
<td>16:9</td>
</tr>
<tr>
<td>SXGA</td>
<td>1280</td>
<td>1024</td>
<td>5:4</td>
</tr>
<tr>
<td>UXGA</td>
<td>1600</td>
<td>1200</td>
<td>4:3</td>
</tr>
<tr>
<td>FHD</td>
<td>1920</td>
<td>1080</td>
<td>16:9</td>
</tr>
<tr>
<td>WUXGA</td>
<td>1920</td>
<td>1200</td>
<td>16:10</td>
</tr>
<tr>
<td>HXGA</td>
<td>4096</td>
<td>3072</td>
<td>4:3</td>
</tr>
<tr>
<td>WQUXGA</td>
<td>3840</td>
<td>2400</td>
<td>16:10</td>
</tr>
<tr>
<td>WHUXGA</td>
<td>7680</td>
<td>4800</td>
<td>16:10</td>
</tr>
</tbody>
</table>
over 700 Mbits/s and 2K digital cinema data exceeds 5 Gbits/sec in uncompressed form. Hence, digital video must be stored and transmitted in compressed form, which leads to compression standards.

Video compression is a key enabling technology for digital video. Standardization of image and video compression is required to ensure compatibility of digital-video products and hardware by different vendors. As a result, several video-compression standards have been developed, and work for even more efficient compression is ongoing. Major standards for image and video compression are listed in Table 2.3.

Historically, standardization in digital-image communication started with the ITU-T (formerly CCITT) digital fax standards. The ITU-T Recommendation T.4 using 1D coding for digital fax transmission was ratified in 1980. Later, a more efficient 2D compression technique was added as an option to the ITU-T recommendation T.30 and ISO JBIG was developed to fix some of the problems with the ITU-T Group 3 and 4 codes, mainly in the transmission of half-tone images.

JPEG was the first color still-image compression standard. It has also found some use in frame-by-frame video compression, called motion JPEG, mostly because of its wide availability in hardware. Later JPEG2000 was developed as a more efficient alternative especially at low bit rates. However, it has mainly found use in the digital cinema standards.

The first commercially successful video-compression standard was MPEG-1 for video storage on CD, which is now obsolete. MPEG-2 was developed for compression of SDTV and HDTV as well as video storage in DVD and was the enabling technology of digital TV. MPEG-4 AVC and HEVC were later developed as more efficient compression standards especially for HDTV and UHDTV as well as video on blu-ray discs. We discuss image- and video-compression technologies and standards in detail in Chapter 7 and Chapter 8, respectively.

<table>
<thead>
<tr>
<th>Standard</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITU-T (formerly CCITT) G3/G4</td>
<td>FAX, Binary images</td>
</tr>
<tr>
<td>ISO JBIG</td>
<td>Binary/halftone, gray-scale images</td>
</tr>
<tr>
<td>ISO JPEG</td>
<td>Still images</td>
</tr>
<tr>
<td>ISO JPEG2000</td>
<td>Digital cinema</td>
</tr>
<tr>
<td>ISO MPEG2</td>
<td>Digital video, SDTV, HDTV</td>
</tr>
<tr>
<td>ISO MPEG4 AVC/ITU-T H.264</td>
<td>Digital video</td>
</tr>
<tr>
<td>ISO HEVC/ITU-T H.265</td>
<td>HD video, HDTV, UHDTV</td>
</tr>
</tbody>
</table>
2.4 3D Video

3D cinema has gained wide acceptance in theatres as many movies are now produced in 3D. Flat-panel 3D TV has also been positively received by consumers for watching sports broadcasts and blu-ray movies. Current 3D video displays are stereoscopic and are viewed by special glasses. Stereo-video formats can be classified as frame-compatible (mainly for broadcast TV) and full-resolution (sequential) formats. Alternatively, multi-view and super multi-view 3D video displays are currently being developed for autostereoscopic viewing. Multi-view video formats without accompanying depth information require extremely high data rates. Multi-view-plus-depth representation and compression are often preferred for efficient storage and transmission of multi-view video as the number of views increases. There are also volumetric, holoscopic (integral imaging), and holographic 3D video formats, which are mostly considered as futuristic at this time.

The main technical obstacles for 3D TV and video to achieve much wider acceptance at home are: i) developing affordable, free-viewing natural 3D display technologies with high spatial, angular, and depth resolution, and ii) capturing and producing 3D content in a format that is suitable for these display technologies. We discuss 3D display technologies and 3D video formats in more detail below.

2.4.1 3D-Display Technologies

A 3D display should ideally reproduce a light field that is an indistinguishable copy of the actual 3D scene. However, this is a rather difficult task to achieve with today's technology due to very large amounts of data that needs to be captured, processed, and stored/transmitted. Hence, current 3D displays can only reproduce a limited set of 3D visual cues instead of the entire light field; namely, they reproduce:

- Binocular depth – Binocular disparity in a stereo pair provides relative depth cue. 3D displays that present only two views, such as stereo TV and digital cinema, can only provide binocular depth cue.
- Head-motion parallax – Viewers expect to see a scene or objects from a slightly different perspective when they move their head. Multi-view, light-field, or volumetric displays can provide head-motion parallax, although most displays can provide only limited parallax, such as only horizontal parallax.

We can broadly classify 3D display technologies as multiple-image (stereoscopic and auto-stereoscopic), light-field, and volumetric displays, as summarized in
Figure 2.10. *Multiple-image displays* present two or more images of a scene by some multiplexing of color sub-pixels on a planar screen such that the right and left eyes see two separate images with binocular disparity, and rely upon the brain to fuse the two images to create the sensation of 3D. *Light-field displays* present light rays as if they are originating from a real 3D object/scene using various technologies such that each pixel of the display can emit multiple light rays with different color, intensity, and directions, as opposed to multiplexing pixels among different views. *Volumetric displays* aim to reconstruct a visual representation of an object/scene using voxels with three physical dimensions via emission, scattering, or relaying of light from a well-defined region in the physical \((x_1, x_2, x_3)\) space, as opposed to displaying light rays emitted from a planar screen.

**Multiple-Image Displays**

Multiple-image displays can be classified as those that require glasses (stereoscopic) and those that don’t (auto-stereoscopic).

Stereoscopic displays present two views with binocular disparity, one for the left and one for the right eye, from a single viewpoint. Glasses are required to ensure that only the right eye sees the right view and the left eye sees the left view. The glasses can be passive or active. Passive glasses are used for color (wavelength) or polarization multiplexing of the two views. Anaglyph is the oldest form of 3D display by color multiplexing using red and cyan filters. Polarization multiplexing applies horizontal and vertical (linear), or clockwise and counterclockwise (circular) polarization to the left and right views, respectively. Glasses apply matching polarization to the right and left eyes. The display shows both left and right views laid over each other with polarization matching that of the glasses in every frame. This will lead to some loss of spatial resolution since half of the sub-pixels in the display panel will be allocated to the left and right views, respectively, using polarized filters. Active glasses (also called active shutter) present the left image to only the left eye by blocking the view of the right eye while the left image is being displayed and vice versa. The display alternates
full-resolution left and right images in sequential order. The active 3D system must assure proper synchronism between the display and glasses. 3D viewing with passive or active glasses is the most developed and commercially available form of 3D display technology. We note that two-view displays lack head-motion parallax and can only provide 3D viewing from a single point of view (from the point where the right and left views have actually been captured) no matter from which angle the viewer looks at the screen. Furthermore, polarization may cause loss of some light due to polarization filter absorption, which may affect scene brightness.

Auto-stereoscopic displays do not require glasses. They can display two views or multiple views. Separation of views can be achieved by different optics technologies, such as parallax barriers or lenticular sheets, so that only certain rays are emitted in certain directions. They can provide head-motion parallax, in addition to binocular depth cues, by either using head-tracking to display two views generated according to head/eye position of the viewer or displaying multiple fixed views. In the former, the need for head-tracking, real-time view generation, and dynamic optics to steer two views in the direction of the viewer gaze increases hardware complexity. In the latter, continuous-motion parallax is not possible with a limited number of views, and proper 3D vision is only possible from some select viewing positions, called sweet spots. In order to determine the number of views, we divide the head-motion range into 2 cm intervals (zones) and present a view for each zone. Then, images seen by the left and right eyes (separated by 6 cm) will be separated by three views. If we allow 4-5 cm head movement toward the left and right, then the viewing range can be covered by a total of eight or nine views. The major drawbacks of autostereoscopic multi-view displays are: i) multiple views are displayed over the same physical screen, sharing sub-pixels between views in a predetermined pattern, which results in loss of spatial resolution; ii) cross-talk between multiple views is unavoidable due to limitations of optics; and iii) there may be noticeable parallax jumps from view to view with a limited number of viewing zones. Due to these reasons, auto-stereoscopic displays have not entered the mass consumer market yet.

State-of-the-art stereoscopic and auto-stereoscopic displays have been reviewed in [Ure 11]. Detailed analysis of stereoscopic and auto-stereoscopic displays from a signal-processing perspective and their quality profiles are provided in [Boe 13].

**Light-Field and Holographic Displays**

Super multi-view (SMV) displays can display up to hundreds of views of a scene taken from different angles (instead of just a right and left view) to create a see-around effect as the viewer slightly changes his/her viewing (gaze) angle. SMV displays employ more advanced optical technologies than just allocating certain
sub-pixels to certain views [Ure 11]. The characteristic parameters of a light-field display are spatial, angular, and perceived depth resolution. If the number of views is sufficiently large such that viewing zones are less than 3 mm, two or more views can be displayed within each eye pupil to overcome the accommodation-vergence conflict and offer a real 3D viewing experience. Quality measures for 3D light-field displays have been studied in [Kov 14].

Holographic imaging requires capturing amplitude (intensity), phase differences (interference pattern), and wavelength (color) of a light field using a coherent light source (laser). Holoscopic imaging (or integral imaging) does not require a coherent light source, but employs an array of microlenses to capture and reproduce a 4D light field, where each lens shows a different view depending on the viewing angle.

**Volumetric Displays**

Different volumetric display technologies aim at creating a 3D viewing experience by means of rendering illumination within a volume that is visible to the unaided eye either directly from the source or via an intermediate surface such as a mirror or glass, which can undergo motion such as oscillation or rotation. They can be broadly classified as swept-volume displays and static volume displays. Swept-volume 3D displays rely on the persistence of human vision to fuse a series of slices of a 3D object, which can be rectangular, disc-shaped, or helical cross-sectioned, into a single 3D image. Static-volume 3D displays partition a finite volume into addressable volume elements, called voxels, made out of active elements that are transparent in “off” state but are either opaque or luminous in “on” state. The resolution of a volumetric display is determined by the number of voxels. It is possible to display scenes with viewing-position-dependent effects (e.g., occlusion) by including transparency (alpha) values for voxels. However, in this case, the scene may look distorted if viewed from positions other than those it was generated for.

The light-field, volumetric, and holographic display technologies are still being developed in major research laboratories around the world and cannot be considered as mature technologies at the time of writing. Note that light-field and volumetric-video representations require orders of magnitude more data (and transmission bandwidth) compared to stereoscopic video. In the following, we cover representations for two-view, multi-view, and super multi-view video.

### 2.4.2 Stereoscopic Video

Stereoscopic two-view video formats can be classified as frame-compatible and full-resolution formats.
2.4 3D Video

Frame-compatible stereo-video formats have been developed to provide 3DTV services over existing digital TV broadcast infrastructures. They employ pixel sub-sampling in order to keep the frame size and rate the same as that of monocular 2D video. Common sub-sampling patterns include side-by-side, top-and-bottom, line interleaved, and checkerboard. Side-by-side format, shown in Figure 2.11(a), applies horizontal subsampling to the left and right views, reducing horizontal resolution by 50%. The subsampled frames are then put together side-by-side. Likewise, top-and-bottom format, shown in Figure 2.11(b), vertically subsamples the left and right views, and stitches them over-under. In the line-interleaved format, the left and right views are again sub-sampled vertically, but put together in an interleaved fashion. Checkerboard format sub-samples left and right views in an offset grid pattern and multiplexes them into a single frame in a checkerboard layout. Among these formats, side-by-side and top-and-bottom are selected as mandatory for broadcast by the latest HDMI specification 1.4a [HDM 13]. Frame-compatible formats are also supported by the stereo and multi-view extensions of the most recent joint MPEG and ITU video-compression standards such as AVC and HEVC (see Chapter 8).

The two-view full resolution stereo is the format of choice for movie and game content. Frame packing, which is a supported format in the HDMI specification version 1.4a, stores frames of left and right views sequentially, without any change in resolution. This full HD stereo-video format requires, in the worst case, twice as much bandwidth as that of monocular video. The extra bandwidth requirement may be kept around 50% by using the Multi-View Video Coding (MVC) standard, which is selected by the Blu-ray Disc Association as the coding format for 3D video.

2.4.3 Multi-View Video

Multi-view and super multi-view displays employ multi-view video representations with varying number of views. Since the required data rate increases linearly with the number of views, depth-based representations are more efficient for multi-view video with more than a few views. Depth-based representations also
enable: i) generation of desired intermediate views that are not present among the original views by using depth-image based rendering (DIBR) techniques, and ii) easy manipulation of depth effects to adjust vergence vs. accommodation conflict for best viewing comfort.

View-plus-depth has initially been proposed as a stereo-video format, where a single view and associated depth map are transmitted to render a stereo pair at the decoder. It is backward compatible with legacy video using a layered bit stream with an encoded view and encoded depth map as a supplementary layer. MPEG specified a container format for view-plus-depth data, called MPEG-C Part 3 [MPG 07], which was later extended to multi-view-video-plus-depth (MVD) format [Smo 11], where $N$ views and $N$ depth maps are encoded and transmitted to generate $M$ views at the decoder, with $N \leq M$. The MVD format is illustrated in Figure 2.12, where only 6 views and 6 depth maps per frame are encoded to reconstruct 45 views per frame at the decoder side by using DIBR techniques.

The depth information needs to be accurately captured/computed, encoded, and transmitted in order to render intermediate views accurately using the received reference view and depth map. Each frame of the depth map conveys the distance of the corresponding video pixel from the camera. Scaled depth values, represented by 8 bits, can be regarded as a separate gray-scale video, which can be compressed very efficiently using state-of-the-art video codecs. Depth map typically requires 15–20%
of the bitrate necessary to encode the original video due to its smooth and less-structured nature.

A difficulty with the view-plus-depth format is generation of accurate depth maps. Although there are time-of-flight cameras that can generate depth or disparity maps, they typically offer limited performance in outdoors environments. Algorithms for depth and disparity estimation by image rectification and disparity matching have been studied in the literature [Kau 07]. Another difficulty is the appearance of regions in the rendered views, which are occluded in the available views. These disocclusion regions may be concealed by smoothing the original depth-map data to avoid appearance of holes. Also, it is possible to use multiple view-plus-depth data to prevent disocclusions [Mul 11]. An extension of the view-plus-depth, which allows better modeling of occlusions, is the layered depth video (LDV). LDV provides multiple depth values for each pixel in a video frame.

While high-definition digital-video products have gained universal user acceptance, there are a number of challenges to overcome in bringing 3D video to consumers. Most importantly, advances in autostereoscopic (without glasses) multi-view display technology will be critical for practical usability and consumer acceptance of 3D viewing technology. Availability of high-quality 3D content at home is another critical factor. In summary, both content creators and display manufacturers need further effort to provide consumers with a high-quality 3D experience without viewing discomfort or fatigue and high transition costs. It seems that the TV/consumer electronics industry has moved its focus to bringing ultra-high-definition products to consumers until there is more progress with these challenges.

2.5 Digital-Video Applications

Main consumer applications for digital video include digital TV broadcasts, digital cinema, video playback from DVD or blu-ray players, as well as video streaming and videoconferencing over the Internet (wired or wireless) [Pit 13].

2.5.1 Digital TV

A digital TV (DTV) broadcasting system consists of video/audio compression, multiplex and transport protocols, channel coding, and modulation subsystems. The biggest single innovation that enabled digital TV services has been advances in video compression since the 1990s. Video-compression standards and algorithms are covered in detail in Chapter 8. Video and audio are compressed separately by different encoders to produce video and audio packetized elementary streams (PES). Video and
audio PES and related data are multiplexed into an MPEG *program stream* (PS). Next, one or more PSs are multiplexed into an MPEG *transport stream* (TS). TS packets are 188-bytes long and are designed with synchronization and recovery in mind for transmission in lossy environments. The TS is then modulated into a signal for transmission. Several different modulation methods exist that are specific to the medium of transmission, which are terrestrial (fixed reception), cable, satellite, and mobile reception.

There are different digital TV broadcasting standards that are deployed globally. Although they all use MPEG-2 or MPEG-4 AVC/H.264 video compression, more or less similar audio coding, and the same transport stream protocol, their channel coding, transmission bandwidth and modulation systems differ slightly. These include the Advanced Television System Committee (ATSC) in the USA, Digital Video Broadcasting (DVB) in Europe, Integrated Multimedia Broadcasting (ISDB) in Japan, and Digital Terrestrial Multimedia Broadcasting in China.

**ATSC Standards**

The first DTV standard was ATSC Standard A/53, which was published in 1995 and was adopted by the Federal Communications Commission in the United States in 1996. This standard supported MPEG-2 Main profile video encoding and 5.1-channel surround sound using Dolby Digital AC-3 encoding, which was standardized as A/52. Support for AVC/H.264 video encoding was added with the ATSC Standard A/72 that was approved in 2008. ATSC signals are designed to use the same 6 MHz bandwidth analog NTSC television channels. Once the digital video and audio signals have been compressed and multiplexed, ATSC uses a 188-byte MPEG transport stream to encapsulate and carry several video and audio programs and metadata. The transport stream is modulated differently depending on the method of transmission:

- Terrestrial broadcasters use 8-VSB modulation that can transmit at a maximum rate of 19.39 Mbit/s. ATSC 8-VSB transmission system adds 20 bytes of Reed-Solomon forward-error correction to create packets that are 208 bytes long.
- Cable television stations operate at a higher signal-to-noise ratio than terrestrial broadcasters and can use either 16-VSB (defined by ATSC) or 256-QAM (defined by Society of Cable Telecommunication Engineers) modulation to achieve a throughput of 38.78 Mbit/s, using the same 6-MHz channel.
- There is also an ATSC standard for satellite transmission; however, direct-broadcast satellite systems in the United States and Canada have long used
either DVB-S (in standard or modified form) or a proprietary system such as DSS (Hughes) or DigiCipher 2 (Motorola).

The receiver must demodulate and apply error correction to the signal. Then, the transport stream may be de-multiplexed into its constituent streams before audio and video decoding.

The newest edition of the standard is ATSC-3.0, which employs the HEVC/H.265 video codec, with OFDM instead of 8-VSB for terrestrial modulation, allowing for 28 Mbps or more of bandwidth on a single 6-MHz channel.

**DVB Standards**

DVB is a suite of standards, adopted by the European Telecommunications Standards Institute (ETSI) and supported by European Broadcasting Union (EBU), which defines the physical layer and data-link layer of the distribution system. The DVB texts are available on the ETSI website. They are specific for each medium of transmission, which we briefly review.

*DVB-T and DVB-T2*

DVB-T is the DVB standard for terrestrial broadcast of digital television and was first published in 1997. It specifies transmission of MPEG transport streams, containing MPEG-2 or H.264/MPEG-4 AVC compressed video, MPEG-2 or Dolby Digital AC-3 audio, and related data, using coded orthogonal frequency-division multiplexing (COFDM) or OFDM modulation. Rather than carrying data on a single radio frequency (RF) channel, COFDM splits the digital data stream into a large number of lower rate streams, each of which digitally modulates a set of closely spaced adjacent sub-carrier frequencies. There are two modes: 2K-mode (1,705 sub-carriers that are 4 kHz apart) and 8K-mode (6,817 sub-carriers that are 1 kHz apart). DVB-T offers three different modulation schemes (QPSK, 16QAM, 64QAM). It was intended for DTV broadcasting using mainly VHF 7 MHz and UHF 8 MHz channels. The first DVB-T broadcast was realized in the UK in 1998. The DVB-T2 is the extension of DVB-T that was published in June 2008. With several technical improvements, it provides a minimum 30% increase in payload, under similar channel conditions compared to DVB-T. The ETSI adopted the DVB-T2 in September 2009.

*DVB-S and DVB-S2*

DVB-S is the original DVB standard for satellite television. Its first release dates back to 1995, while development lasted until 1997. The standard only specifies physical...
link characteristics and framing for delivery of MPEG transport stream (MPEG-TS) containing MPEG-2 compressed video, MPEG-2 or Dolby Digital AC-3 audio, and related data. The first commercial application was in Australia, enabling digitally broadcast, satellite-delivered television to the public. DVB-S has been used in both multiple-channel per carrier and single-channel per carrier modes for broadcast network feeds and direct broadcast satellite services in every continent of the world, including Europe, the United States, and Canada.

DVB-S2 is the successor of the DVB-S standard. It was developed in 2003 and ratified by the ETSI in March 2005. DVB-S2 supports broadcast services including standard and HDTV, interactive services including Internet access, and professional data content distribution. The development of DVB-S2 coincided with the introduction of HDTV and H.264 (MPEG-4 AVC) video codecs. Two new key features that were added compared to the DVB-S standard are:

- A powerful coding scheme, Irregular Repeat-Accumulate codes, based on a modern LDPC code, with a special structure for low encoding complexity.
- Variable coding and modulation (VCM) and adaptive coding and modulation (ACM) modes to optimize bandwidth utilization by dynamically changing transmission parameters.

Other features include enhanced modulation schemes up to 32-APSK, additional code rates, and introduction of a generic transport mechanism for IP packet data including MPEG-4 AVC video and audio streams, while supporting backward compatibility with existing DVB-S transmission. The measured DVB-S2 performance gain over DVB-S is around a 30% increase of available bitrate at the same satellite transponder bandwidth and emitted signal power. With improvements in video compression, an MPEG-4 AVC HDTV service can now be delivered in the same bandwidth used for an early DVB-S based MPEG-2 SDTV service. In March 2014, the DVB-S2X specification was published as an optional extension adding further improvements.

**DVB-C and DVB-C2**

The DVB-C standard is for broadcast transmission of digital television over cable. This system transmits an MPEG-2 or MPEG-4 family of digital audio/digital video stream using QAM modulation with channel coding. The standard was first published by the ETSI in 1994, and became the most widely used transmission system for digital cable television in Europe. It is deployed worldwide in systems ranging
2.5 Digital-Video Applications

from larger cable television networks (CATV) to smaller satellite master antenna TV (SMATV) systems.

The second-generation DVB cable transmission system DVB-C2 specification was approved in April 2009. DVB-C2 allows bitrates up to 83.1 Mbit/s on an 8 MHz channel when using 4096-QAM modulation, and up to 97 Mbit/s and 110.8 Mbit/s per channel when using 16384-QAM and 65536-AQAM modulation, respectively. By using state-of-the-art coding and modulation techniques, DVB-C2 offers more than a 30% higher spectrum efficiency under the same conditions, and the gains in downstream channel capacity are greater than 60% for optimized HFC networks. These results show that the performance of the DVB-C2 system gets so close to the theoretical Shannon limit that any further improvements would most likely not be able to justify the introduction of a disruptive third generation cable-transmission system.

There is also a DVB-H standard for terrestrial mobile TV broadcasting to hand-held devices. The competitors of this technology have been the 3G cellular-system-based MBMS mobile-TV standard, the ATSC-M/H format in the United States, and the Qualcomm MediaFLO. DVB-SH (satellite to handhelds) and DVB-NGH (Next Generation Handheld) are possible future enhancements to DVB-H. However, none of these technologies have been commercially successful.

2.5.2 Digital Cinema

Digital cinema refers to digital distribution and projection of motion pictures as opposed to use of motion picture film. A digital cinema theatre requires a digital projector (instead of a conventional film projector) and a special computer server. Movies are supplied to theatres as digital files, called a Digital Cinema Package (DCP), whose size is between 90 gigabytes (GB) and 300 GB for a typical feature movie. The DCP may be physically delivered on a hard drive or can be downloaded via satellite. The encrypted DCP file first needs to be copied onto the server. The decryption keys, which expire at the end of the agreed upon screening period, are supplied separately by the distributor. The keys are locked to the server and projector that will screen the film; hence, a new set of keys are required to show the movie on another screen. The playback of the content is controlled by the server using a playlist.

Technology and Standards

Digital cinema projection was first demonstrated in the United States in October 1998 using Texas Instruments’ DLP projection technology. In January 2000, the
Society of Motion Picture and Television Engineers, in North America, initiated a group to develop digital cinema standards. The Digital Cinema Initiative (DCI), a joint venture of six major studios, was established in March 2002 to develop a system specification for digital cinema to provide robust intellectual property protection for content providers. DCI published the first version of a specification for digital cinema in July 2005. Any DCI-compliant content can play on any DCI-compliant hardware anywhere in the world.

Digital cinema uses high-definition video standards, aspect ratios, or frame rates that are slightly different than HDTV and UHDTV. The DCI specification supports 2K (2048 × 1080 or 2.2 Mpixels) at 24 or 48 frames/sec and 4K (4096 × 2160 or 8.8 Mpixels) at 24 frames/sec modes, where resolutions are represented by the horizontal pixel count. The 48 frames/sec is called high frame rate (HFR). The specification employs the ISO/IEC 15444-1 JPEG2000 standard for picture encoding, and the CIE XYZ color space is used at 12 bits per component encoded with a 2.6 gamma applied at projection. It ensures that 2K content can play on 4K projectors and vice versa.

Digital Cinema Projectors

Digital cinema projectors are similar in principle to other digital projectors used in the industry. However, they must be approved by the DCI for compliance with the DCI specifications: i) they must conform to the strict performance requirements, and ii) they must incorporate anti-piracy protection to protect copyrights. Major DCI-approved digital cinema projector manufacturers include Christie, Barco, NEC, and Sony. The first three manufacturers have licensed the DLP technology from Texas Instruments, and Sony uses its own SXRD technology. DLP projectors were initially available in 2K mode only. DLP projectors became available in both 2K and 4K in early 2012, when Texas Instruments’ 4K DLP chip was launched. Sony SXRD projectors are only manufactured in 4K mode.

DLP technology is based on digital micromirror devices (DMDs), which are chips whose surface is covered by a large number of microscopic mirrors, one for each pixel; hence, a 2K chip has about 2.2 million mirrors and a 4K chip about 8.8 million. Each mirror vibrates several thousand times a second between on and off positions. The proportion of the time the mirror is in each position varies according to the brightness of each pixel. Three DMD devices are used for color projection, one for each of the primary colors. Light from a Xenon lamp, with power between 1 kW and 7 kW, is split by color filters into red, green, and blue beams that are directed at the appropriate DMD.
Transition to digital projection in cinemas is ongoing worldwide. According to the National Association of Theatre Owners, 37,711 screens out of 40,048 in the United States had been converted to digital and about 15,000 were 3D capable as of May 2014.

3D Digital Cinema

The number of 3D-capable digital cinema theatres is increasing with wide interest of audiences in 3D movies and an increasing number of 3D productions. A 3D-capable digital cinema video projector projects right-eye and left-eye frames sequentially. The source video is produced at 24 frames/sec per eye; hence, a total of 48 frames/sec for right and left eyes. Each frame is projected three times to reduce flicker, called triple flash, for a total of 144 times per second. A silver screen is used to maintain light polarization upon reflection. There are two types of stereoscopic 3D viewing technology where each eye sees only its designated frame: i) glasses with polarizing filters oriented to match projector filters, and ii) glasses with liquid crystal (LCD) shutters that block or transmit light in sync with the projectors. These technologies are provided under the brands RealD, MasterImage, Dolby 3D, and XpanD.

The polarization technology combines a single 144-Hz digital projector with either a polarizing filter (for use with polarized glasses and silver screens) or a filter wheel. RealD 3D cinema technology places a push-pull electro-optical liquid crystal modulator called a ZScreen in front of the projector lens to alternately polarize each frame. It circularly polarizes frames clockwise for the right eye and counter-clockwise for the left eye. MasterImage uses a filter wheel that changes the polarity of the projector’s light output several times per second to alternate the left-and-right-eye views. Dolby 3D also uses a filter wheel. The wheel changes the wavelengths of colors being displayed, and tinted glasses filter these changes so the incorrect wavelength cannot enter the wrong eye. The advantage of circular polarization over linear polarization is that viewers are able to slightly tilt their head without seeing double or darkened images.

The XpanD system alternately flashes the images for each eye that viewers observe using electronically synchronized glasses. The viewer wears electronic glasses whose LCD lenses alternate between clear and opaque to show only the correct image at the correct time for each eye. XpanD uses an external emitter that broadcasts an invisible infrared signal in the auditorium that is picked up by glasses to synchronize the shutter effect.

IMAX Digital 3D uses two separate 2K projectors that represent the left and right eyes. They are separated by a distance of 64 mm (2.5 in), which is the average distance
between a human’s eyes. The two 2K images are projected over each other (superposed) on a silver screen with proper polarization, which makes the image brighter. Right and left frames on the screen are directed only to the correct eye by means of polarized glasses that enable the viewer to see in 3D. Note that IMAX theatres use the original 15/70 IMAX higher resolution frame format on larger screens.

2.5.3 Video Streaming over the Internet

Video streaming refers to delivery of media over the Internet, where the client player can begin playback before the entire file has been sent by the server. A server-client streaming system consists of a streaming server and a client that communicate using a set of standard protocols. The client may be a standalone player or a plugin as part of a Web browser. The streaming session can be a video-on-demand request (sometimes called a pull-application) or live Internet broadcasting (called a push-application). In a video-on-demand session, the server streams from a pre-encoded and stored file. Live streaming refers to live content delivered in real-time over the Internet, which requires a live camera and a real-time encoder on the server side.

Since the Internet is a best-effort channel, packets may be delayed or dropped by the routers and the effective end-to-end bitrates fluctuate in time. Adaptive streaming technologies aim to adapt the video-source (encoding) rate according to an estimate of the available end-to-end network rate. One possible way to do this is stream switching, where the server encodes source video at multiple pre-selected bitrates and the client requests switching to the stream encoded at the rate that is closest to its network access rate. A less commonly deployed solution is based on scalable video coding, where one or more enhancement layers of video may be dropped to reduce the bitrate as needed.

In the server-client model, the server sends a different stream to each client. This model is not scalable, since server load increases linearly with the number of stream requests. Two solutions to solve this problem are multicasting and peer-to-peer (P2P) streaming. We discuss the server-client, multicast, and P2P streaming models in more detail below.

Server-Client Streaming

This is the most commonly used streaming model on the Internet today. All video streaming systems deliver video and audio streams by using a streaming protocol built on top of transmission control protocol (TCP) or user datagram protocol (UDP). Streaming solutions may be based on open-standard protocols published by
the Internet Engineering Task Force (IETF) such as RTP/UDP or HTTP/TCP, or may be proprietary systems, where RTP stands for real-time transport protocol and HTTP stands for hyper-text transfer protocol.

**Streaming Protocols**

Two popular streaming protocols are Real-Time Streaming Protocol (RTSP), an open standard developed and published by the IETF as RFC 2326 in 1998, and Real Time Messaging Protocol (RTMP), a proprietary solution developed by Adobe Systems.

RTSP servers use the Real-time Transport Protocol (RTP) for media stream delivery, which supports a range of media formats (such as AVC/H.264, MJPEG, etc.). Client applications include QuickTime, Skype, and Windows Media Player. Android smartphone platforms also include support for RTSP as part of the 3GPP standard.

RTMP is primarily used to stream audio and video to Adobe's Flash Player client. The majority of streaming videos on the Internet is currently delivered via RTMP or one of its variants due to the success of the Flash Player. RTMP has been released for public use. Adobe has included support for adaptive streaming into the RTMP protocol.

The main problem with UDP-based streaming is that streams are frequently blocked by firewalls, since they are not being sent over HTTP (port 80). In order to circumvent this problem, protocols have been extended to allow for a stream to be encapsulated within HTTP requests, which is called tunneling. However, tunneling comes at a performance cost and is often only deployed as a fallback solution. Streaming protocols also have secure variants that use encryption to protect the stream.

**HTTP Streaming**

Streaming over HTTP, which is a more recent technology, works by breaking a stream into a sequence of small HTTP-based file downloads, where each download loads one short *chunk* of the whole stream. All flavors of HTTP streaming include support for adaptive streaming (bitrate switching), which allows clients to dynamically switch between different streams of varying quality and chunk size during playback, in order to adapt to changing network conditions and available CPU resources. By using HTTP, firewall issues are generally avoided. Another advantage of HTTP streaming is that it allows HTTP chunks to be cached within ISPs or
corporations, which would reduce the bandwidth required to deliver HTTP streams, in contrast to video streamed via RTMP.

Different vendors have implemented different HTTP-based streaming solutions, which all use similar mechanisms but are incompatible; hence, they all require the vendor’s own software:

- HTTP Live Streaming (HLS) by Apple is an HTTP-based media streaming protocol that can dynamically adjust movie playback quality to match the available speed of wired or wireless networks. HTTP Live Streaming can deliver streaming media to an iOS app or HTML5-based website. It is available as an IETF Draft (as of October 2014) [Pan 14].
- Smooth Streaming by Microsoft enables adaptive streaming of media to clients over HTTP. The format specification is based on the ISO base media file format. Microsoft provides Smooth Streaming Client software development kits for Silverlight and Windows Phone 7.
- HTTP Dynamic Streaming (HDS) by Adobe provides HTTP-based adaptive streaming of high-quality AVC/H.264 or VP6 video for a Flash Player client platform.

MPEG-DASH is the first adaptive bit-rate HTTP-based streaming solution that is an international standard, published in April 2012. MPEG-DASH is audio/video codec agnostic. It allows devices such as Internet-connected televisions, TV set-top boxes, desktop computers, smartphones, tablets, etc., to consume multimedia delivered via the Internet using previously existing HTTP web server infrastructure, with the help of adaptive streaming technology. Standardizing an adaptive streaming solution aims to provide confidence that the solution can be adopted for universal deployment, compared to similar proprietary solutions such as HLS by Apple, Smooth Streaming by Microsoft, or HDS by Adobe. An implementation of MPEG-DASH using a content centric networking (CCN) naming scheme to identify content segments is publicly available [Led 13]. Several issues still need to be resolved, including legal patent claims, before DASH can become a widely used standard.

Multicast and Peer-to-Peer (P2P) Streaming
Multicast is a one-to-many delivery system, where the source server sends each packet only once, and the nodes in the network replicate packets only when necessary to reach multiple clients. The client nodes send join and leave messages, e.g., as in the
2.5 Digital-Video Applications

case of Internet television when the user changes the TV channel. In P2P streaming, clients (peers) forward packets to other peers (as opposed to network nodes) to minimize the load on the source server.

The multicast concept can be implemented at the IP or application level. The most common transport layer protocol to use multicast addressing is the User Datagram Protocol (UDP). IP multicast is implemented at the IP routing level, where routers create optimal distribution paths for datagrams sent to a multicast destination address. IP multicast has been deployed in enterprise networks and multimedia content delivery networks, e.g., in IPTV applications. However, IP multicast is not implemented in commercial Internet backbones mainly due to economic reasons. Instead, application layer multicast-over-unicast overlay services for application-level group communication are widely used.

In media streaming over P2P overlay networks, each peer forwards packets to other peers in a live media streaming session to minimize the load on the server. Several protocols that help peers find a relay peer for a specified stream exist [Gu14]. There are P2PTV networks based on real-time versions of the popular file-sharing protocol BitTorrent. Some P2P technologies employ the multicast concept when distributing content to multiple recipients, which is known as peercasting.

2.5.4 Computer Vision and Scene/Activity Understanding

Computer vision is a discipline of computer science that aims to duplicate abilities of human vision by processing and understanding digital images and video. It is such a large field that it is the subject of many excellent textbooks [Har04, For11, Sze11]. The visual data to be processed can be still images, video sequences, or views from multiple cameras. Computer vision is generally divided into high-level and low-level vision. High-level vision is often considered as part of artificial intelligence and is concerned with the theory of learning and pattern recognition with application to object/activity recognition in order to extract information from images and video. We mention computer vision here because many of the problems addressed in image/video processing and low-level vision are common. Low-level vision includes many image- and video-processing tasks that are the subject of this book such as edge detection, image enhancement and restoration, motion estimation, 3D scene reconstruction, image segmentation, and video tracking. These low-level vision tasks have been used in many computer-vision applications, including road monitoring, military surveillance, and robot navigation. Indeed, several of the methods discussed in this book have been developed by computer-vision researchers.
2.6 Image and Video Quality

Video quality may be measured by the quality of experience of viewers, which can usually be reliably measured by subjective methods. There have been many studies to develop objective measures of video quality that correlate well with subjective evaluation results [Cho 14, Bov 13]. However, this is still an active research area. Since analog video is becoming obsolete, we start by defining some visual artifacts related to digital video that are the main cause of loss of quality of experience.

2.6.1 Visual Artifacts

Artifacts are visible distortions in images/videos. We can classify visual artifacts as spatial and temporal artifacts. Spatial artifacts, such as blur, noise, ringing, and blocking, are most disturbing in still images but may also be visible in video. In addition, in video, temporal freeze and skipped frames are important causes of visual disturbance and, hence, loss of quality of experience.

Blur refers to lack or loss of image sharpness (high spatial frequencies). The main causes of blur are insufficient spatial resolution, defocus, and/or motion between camera and the subject. According to the Nyquist sampling theorem, the highest horizontal and vertical spatial frequencies that can be represented is determined by the sampling rate (pixels/cm), which relates to image resolution. Consequently, low-resolution images cannot contain high spatial frequencies and appear blurred. Defocus blur is due to incorrect focus of the camera, which may be due to depth of field. Motion blur is caused by relative movement of the subject and camera while the shutter is open. It may be more noticeable in imaging darker scenes since the shutter has to remain open for longer time.

Image noise refers to low amplitude, high-frequency random fluctuations in the pixel values of recorded images. It is an undesirable by-product of image capture, which can be produced by film grain, photo-electric sensors, and digital camera circuitry, or image compression. It is measured by signal-to-noise ratio. Noise due to electronic fluctuations can be modeled by a white, Gaussian random field, while noise due to LCD sensor imperfections is usually modeled as impulsive (salt-and-pepper) noise. Noise at low-light (signal) levels can be modeled as speckle noise.

Image/video compression also generates noise, known as quantization noise and mosquito noise. Quantization or truncation of the DCT/wavelet transform coefficients results in quantization noise. Mosquito noise is temporal noise, i.e., flickering-like luminance/chrominance fluctuations as a consequence of differences in coding observed in smoothly textured regions or around high contrast edges in consecutive frames of video.
Ringing and blocking artifacts, which are by-products of DCT image/video compression, are also observed in compressed images/video. Ringing refers to oscillations around sharp edges. It is caused by sudden truncation of DCT coefficients due to coarse quantization (also known as the Gibbs effect). DCT is usually taken over $8 \times 8$ blocks. Coarse quantization of DC coefficients may cause mismatch of image mean over $8 \times 8$ blocks, which results in visible block boundaries known as blocking artifacts.

Skip frame and freeze frame are the result of video transmission over unreliable channels. They are caused by video packets that are not delivered on time. When video packets are late, there are two options: skip late packets and continue with the next packet, which is delivered on time, or wait (freeze) until the late packets arrive. Skipped frames result in motion jerkiness and discontinuity, while freeze frame refers to complete stopping of action until the video is rebuffered.

Visibility of artifacts is affected by the viewing conditions, as well as the type of image/video content as a result of spatial and temporal-masking effects. For example, spatial-image artifacts that are not visible in full-motion video may be highly objectionable when we freeze frame.

### 2.6.2 Subjective Quality Assessment

Measurement of subjective video quality can be challenging because many parameters of set-up and viewing conditions, such as room illumination, display type, brightness, contrast, resolution, viewing distance, and the age and educational level of experts, can influence the results. The selection of video content and the duration also affect the results. A typical subjective video quality evaluation procedure consists of the following steps:

1. Choose video sequences for testing
2. Choose the test set-up and settings of system to evaluate
3. Choose a test method (how sequences are presented to experts and how their opinion is collected: DSIS, DSCQS, SSCQE, DSCS)
4. Invite sufficient number and types of experts (18 or more is recommended)
5. Carry out testing and calculate the mean expert opinion scores (MOS) for each test set-up

In order to establish meaningful subjective assessment results, some test methods, grading scales, and viewing conditions have been standardized by ITU-T Recommendation BT.500-11 (2002) “Methodology for the subjective assessment of the quality of television pictures.” Some of these test methods are double stimulus where
viewers rate the quality or change in quality between two video streams (reference and impaired). Others are single stimulus where viewers rate the quality of just one video stream (the impaired). Examples of the former are the double stimulus impairment scale (DSIS), double stimulus continuous quality scale (DSCQS), and double stimulus comparison scale (DSCS) methods. An example of the latter is the single stimulus continuous quality evaluation (SSCQE) method. In the DSIS method, observers are first presented with an unimpaired reference video, then the same video impaired, and he/she is asked to vote on the second video using an impairment scale (from “impairments are imperceptible” to “impairments are very annoying”). In the DSCQS method, the sequences are again presented in pairs: the reference and impaired. However, observers are not told which one is the reference and are asked to assess the quality of both. In the series of tests, the position of the reference is changed randomly. Different test methodologies have claimed advantages for different cases.

2.6.3 Objective Quality Assessment

The goal of objective image quality assessment is to develop quantitative measures that can automatically predict perceived image quality [Bov 13]. Objective image/video quality metrics are mathematical models or equations whose results are expected to correlate well with subjective assessments. The goodness of an objective video-quality metric can be assessed by computing the correlation between the objective scores and the subjective test results. The most frequently used correlation coefficients are the Pearson linear correlation coefficient, Spearman rank-order correlation coefficient, kurtosis, and the outliers ratio.

Objective metrics are classified as full reference (FR), reduced reference (RR), and no-reference (NR) metrics, based on availability of the original (high-quality) video, which is called the reference. FR metrics compute a function of the difference between every pixel in each frame of the test video and its corresponding pixel in the reference video. They cannot be used to evaluate the quality of the received video, since a reference video is not available at the receiver end. RR metrics extract some features of both videos and compare them to give a quality score. Only some features of the reference video must be sent along with the compressed video in order to evaluate the received video quality at the receiver end. NR metrics assess the quality of a test video without any reference to the original video.

Objective Image/Video Quality Measures

Perhaps the most well-established methodology for FR objective image and video quality evaluation is pixel-by-pixel comparison of image/video with the reference.
The peak signal-to-noise ratio (PSNR) measures the logarithm of the ratio of the maximum signal power to the mean square difference (MSE), given by

\[
PSNR = 10 \log_{10} \left( \frac{255^2}{MSE} \right)
\]

where the MSE between the test video \( s[n_1, n_2, k] \), which is \( N_1 \times N_2 \) pixels and \( N_3 \) frames long, and reference video \( s[n_1, n_2, k] \) with the same size, can be computed by

\[
MSE = \frac{1}{N_1 \times N_2 \times N_3} \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} \sum_{k=0}^{N_3-1} (s[n_1, n_2, k] - \hat{s}[n_1, n_2, k])^2
\]

Some have claimed that PSNR may not correlate well with the perceived visual quality since it does not take into account many characteristics of the human visual system, such as spatial- and temporal-masking effects. To this effect, many alternative FR metrics have been proposed. They can be classified as those based on structural similarity and those based on human vision models.

The structural similarity index (SSIM) is a structural image similarity based FR metric that aims to measure perceived change in structural information between two \( N \times N \) luminance blocks \( x \) and \( y \), with means \( \mu_x \) and \( \mu_y \) and variances \( \sigma_x^2 \) and \( \sigma_y^2 \), respectively. It is given by [Wan 04]

\[
SSIM(x, y) = \frac{(2\mu_x \mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}
\]

where \( \sigma_{xy} \) is the covariance between windows \( x \) and \( y \) and \( c_1 \) and \( c_2 \) are small constants to avoid division by very small numbers.

Perceptual evaluation of video quality (PEVQ) is a vision-model-based FR metric that analyzes pictures pixel-by-pixel after a temporal alignment (registration) of corresponding frames of reference and test video. PEVQ aims to reflect how human viewers would evaluate video quality based on subjective comparison and outputs mean opinion scores (MOS) in the range from 1 (bad) to 5 (excellent).

VQM is an RR metric that is based on a general model and associated calibration techniques and provides estimates of the overall impressions of subjective video quality [Pin 04]. It combines perceptual effects of video artifacts including blur, noise, blockiness, color distortions, and motion jerkiness into a single metric.

NR metrics can be used for monitoring quality of compressed images/video or video streaming over the Internet. Specific NR metrics have been developed for
quantifying such image artifacts as noise, blockiness, and ringing. However, the ability of these metrics to make accurate quality predictions are usually satisfactory only in a limited scope, such as for JPEG/JPEG2000 images.

The International Telecommunications Union (ITU) Video Quality Experts Group (VQEG) standardized some of these metrics, including the PEVQ, SSIM, and VQM, as ITU-T Rec. J.246 (RR) and J.247 (FR) in 2008 and ITU-T Rec. J.341 (FR HD) in 2011. It is perhaps useful to distinguish the performance of these structural similarity and human vision model based metrics on still images and video. It is fair to say these metrics have so far been more successful on still images than video for objective quality assessment.

Objective Quality Measures for Stereoscopic 3D Video

FR metrics for evaluation of 3D image/video quality is technically not possible, since the 3D signal is formed only in the brain. Hence, objective measures based on a stereo pair or video-plus-depth-maps should be considered as RR metrics. It is generally agreed upon that 3D quality of experience is related to at least three factors:

- Quality of display technology (cross-talk)
- Quality of content (visual discomfort due to accomodation-vergence conflict)
- Encoding/transmission distortions/ artifacts

In addition to those artifacts discussed in Section 2.6.1, the main factors in 3D video quality of experience are visual discomfort and depth perception. As discussed in Section 2.1.4, visual discomfort is mainly due to the conflict between accommodation and vergence and cross-talk between the left and right views. Human perception of distortions/artifacts in 3D stereo viewing is not fully understood yet. There have been some preliminary works on quantifying visual comfort and depth perception [Uka 08, Sha 13]. An overview of evaluation of stereo and multi-view image/video quality can be found in [Win 13]. There are also some studies evaluating the perceptual quality of symmetrically and asymmetrically encoded stereoscopic videos [Sil 13].

References

References


Index

Numbers
1D-RLC (run-length coding), 419–421
2D
convolution summation. See Convolution summation, 2D
image-plane motion. See Motion estimation
mesh tracking, 327–328
notation, 2
rectangular sampling, 31–32, 37–41
sampling lattices, 32–33
2D apparent-motion estimation
dense-motion estimation, 215–216
displaced-frame difference, 219–220
hierarchical motion estimation, 223–224
as ill-posed problem, 220–223
image registration, 217
optical flow equation/normal flow, 217–219
overview of, 214
performance measures for, 224–225
sparse-correspondence estimation, 214
2D apparent-motion models
defined, 210–214
non-parametric models, 213–214
parametric models, 210–213
2D-AR model, 29
2D-DCT (discrete cosine transform)
hybrid DPCM coding, 464
MC-transform coding, 466

overview of, 18–19
relationship to DFT, 20
2D-DFT (discrete Fourier transform)
boundary effects, 250
computation of, 15
DFT domain implementation, 172
diagonalizing block-Toeplitz matrices, 155
drawbacks of inverse filtering, 170
multi-frame Wiener restoration, 376
properties of, 16
2D-RLC (run-length coding), 419, 421–423
2-tap Haar filters, 445
3D
digital cinema, 91–92
motion. See Motion estimation
steering kernel regression, SR, 394
Taylor series, 394
3D motion/structure estimation
affine reconstruction, 253–255
camera calibration, 252–253
dense structure from zero, 263
Euclidean reconstruction, 260
overview of, 251–252
planar-parallax/relative affine reconstruction,
261–263
projective reconstruction, 255–260
3D scenes, projecting onto 2D image plane,
196–199

557
3D video
- challenges of, 85
- defined, 1
- disparity processing, 63
- display technologies, 79–82
- multi-view, 83–85
- objective quality metrics for stereoscopic, 100
- overview of, 79
- stereoscopic video, 82–83
- 3D-AVC coding, 508–510
- 3D-DCT coding, 463–464
- 3D-HEVC tools, 510–512
- 3D-motion/pose estimation, 250
- 3D-transform coding, 463–466
- 3DTV services, 83
- 3D-wavelet/sub-band coding, 464–466
- 4 × 4 prediction, H.264/AVC, 486
- 16 × 8 prediction for field pictures, MPEG-2 video, 479
- 16 × 16 prediction, H.264/AVC, 486
- 24 Hz movies to 50/60 Hz conversion, 361–363
- 50 to 60 Hz conversion, 363

A
- a posteriori probability. See MAP (maximum a posteriori) probability estimates
- AAC Audio, MPEG-2, 476
- Above-right predictor (ARP), VSBM, 240
- AC coefficients
  - JPEG, 435, 439–440
  - MPEG-1, 471–472
  - MPEG-1 vs. MPEG-2, 481
- Accommodation distance, human stereo vision, 62–63
- Active (active shutter) glasses, 80–81
- Active-contour models (snakes), 287–289
- Active-contour motion tracking, 325–327, 329
- A/D (analog-to-digital) conversion process, 66–67
- Adaptive arithmetic coding, 416, 423–424
- Adaptive filters
  - in image enhancement, 145–146
  - in interpolation, 113
  - LMMSE, 155–157
- Adaptive luminance compensation (ALC), 3D-AVV, 509
- Adaptive MAP method, image segmentation, 284–285
- Adaptive reference filtering coding tool, MVC, 506
- Adaptive smoothness constraints, 232–233
- Adaptive streaming
  - in Adobe Flash with RTMP, 93
  - in HTTP streaming, 93
  - in MPEG-DASH, 94
  - in Smooth Streaming, 94
  - in video streaming over Internet, 82
- Adaptive thresholds
  - in change detection, 293, 295
  - computing, 276
  - in wavelet shrinkage, 160–161
- Adaptive/nonlinear interpolation, 118–119
- Adaptive-weighted-averaging (AWA) filter, 372–373
- Additive color model, 56–57
- Additive noise model, 148
- Advanced residual prediction (ARP), 3D-HEVC, 511
- Advanced Television System Committee (ATSC) standards, 86–87
- Advanced-motion vector prediction (AMVP), HEVC, 496
- AE (angular error), motion estimation, 223–224
- Affine camera
  - in affine reconstruction, 253
  - orthographic projection, 199–200
  - paraperspective projection, 201
  - weak-perspective projection, 201
- Affine model
  - in 2D-mesh tracking, 327–328
  - in active-contour tracking, 325
  - in clustering within motion-parameter, 202–203
- Lukas–Kanade solution for, 229
  - as parametric apparent 2D-motion model, 211–212
Index

Affine reconstruction, 253–255
Affinity-based image matting, 329
AGC (automatic gain control), 139
Aggregation, BM3D filtering, 164
ALC (adaptive luminance compensation), 3D-AVV, 509
Alias-cancellation, analysis-synthesis filters, 123, 445
Aliasing
  avoiding in sampling structure conversion, 45
  DFT and, 14
  IIR filters in DFT domain and, 28
  image decimation and, 111, 112
  in LR images for SR reconstruction, 386–388
  in LR images for super-resolution, 379–380
  from violating Nyquist sampling rate, 39–40
Alpha-trimmed mean filters, image denoising, 159–160
Alternate scan, MPEG-2 video, 480–481
AMVP (Advanced-motion vector prediction), HEVC, 496
Anaglyph, 80
Analog MD signal, 2
Analog video
  3D sampling lattices, 33–34
  analog-to-digital conversion, 33, 66–67
  orthogonal sampling for progressive, 32
  overview of, 63–64
  progressive vs. interlaced scanning, 64–65
  signal formats, 65–66
Analog-to-digital (A/D) conversion process, 66–67
Analog-video signal, 64–66
Analysis filters, 122–124, 444–445
Analysis-synthesis filters, 123, 445
Anchor pictures, MVC standard, 505
Angular error (AE), motion estimation, 223–224
Anisotropic diffusion filters, 157
Anisotropic Gaussian filtering, 119
Anti-alias filtering
  down-conversion with, 351–352
  down-conversion without, 352–353
  in image decimation, 112–113
Aperture problem, 222–223
Apparent motion model, 206–207
Apparent-motion estimation. See 2D apparent-motion estimation
Apparent-motion models, 2D
  non-parametric models, 213–214
  parametric models, 210–213
Arbitrary camera pose, perspective projection, 198–199
Arbitrary slice ordering (ASO), H.264/AVC, 490
Arbitrary-motion trajectories, MC filtering, 345–346
Arithmetic coding
  adaptive, 416
  as entropy coding, 410
  in image compression, 414–417
  JBIG and adaptive, 423–424
ARP (above-right predictor), VSBM, 240
ARP (advanced residual prediction), 3D-HEVC, 511
Artifacts
  aliasing, 116, 120
  analog signal and, 64
  bi-lateral filtering overcoming, 109
  compression, 289, 442–443
  contouring, 71
  human visual system and, 58
  interpolation, 118
  from phase distortions in filtering, 24
  regularization, 171
  spatial masking and, 59
  spatial resolution/frame rate and, 68
  visual digital video, 96–97, 99–100
ASO (arbitrary slice ordering), H.264/AVC, 490
Aspect ratio
  digital video 16:9/16:10, 76
  video signal, 65
Asymmetric coding of stereo video, 506–507
Asymmetric property, DFT, 16
Asymmetric stereo-video encoding, 63
ATSC (Advanced Television System Committee) standards, 86–87
Audio, MPEG-2, 476
Auto-calibration, camera, 252, 260
Automatic gain control (AGC), 139
Auto-stereoscopic (no glasses), 80–81, 85
Auto-stereoscopic displays, cross-talk in, 63
Average central difference, estimating partials, 129
AWA (adaptive-weighted-averaging) filter, 372–373

B
Background modeling, adaptive, 293–295
Background subtraction, change detection
adaptive background modeling, 293–295 exercises, 338
frame differencing methods, 291–293
other approaches, 297–298
overview of, 291
spatial and temporal consistency, 295
ViBe algorithm, 295–297
Backward extension of MVs, MC de-interlacing, 360
Backward-motion estimation, dense-
correspondence, 216
Bandlimited
  circularly, 37–38
  continuous signal as, 39–40
digitizing image that is not, 50
  ideal spatio-temporal interpolation filter, 42
Bandpass spectral response, 54, 62
Bandwidth
  analog TV and, 65
digital TV and, 86–89
  HDMI full stereo-video format and, 83 in HTTP streaming, 94
 Bit rate/quality tradeoff, JPEG, 435
 Bit-depth, digital images/videos, 70–71
  Bitmap parameters, 68–69
  Bit-plane coding, 418–419
  Bit-rates, asymmetric coding of stereo video, 506–507
  Bi-directional prediction, MPEG-1, 472
  Bi-lateral filters, 109–110, 161–162
  Binarization, thresholding for, 276
  Binocular depth cues, 81
  Binocular disparity, 62, 80
  Binocular rivalry, 63
  Bi-orthogonal filters, 125–127, 446
  Bit rate/quality tradeoff, JPEG, 435
  Bit-depth, digital images/videos, 70–71
  Bitmap parameters, 68–69
  Bit-plane coding, 418–419
  Bit-rates, asymmetric coding of stereo video, 506–507
  Bit-stream extraction, SNR scalability, 502
  BitTorrent protocol, P2PTV, 95
  BLA (broken-link access) picture, HEVC, 492
Blind-image restoration, 168, 175–176
Block coding
  JPEG2000, 448, 453–454
  Lempel–Ziv, 430–431
  overview of, 414
Block size
  in basic block-matching, 134–135
  in maximum-displacement estimate, 150
  in motion-compensated prediction, 486–487, 492
  in transform coding, 433
  in variable-size block matching, 138–139
Block translation model, 211, 227
Blocking artifacts, 97, 464, 490
Block-matching, in motion estimation
   basic procedure, 234–238
   as deterministic non-parametric model, 213–214
   fast search in, 236–238
   full search in, 235–236
   generalized, 241–242
   hierarchical, 240–241
   introduction to, 233–234
   sub-pixel search in, 238
   variable-size, 238–240
Blocks, MPEG-1, 470, 472
Block-Toeplitz matrices
   in image denoising, 155
   in image filtering, 168, 170
   in image restoration, 167
   in video filtering, 376
Block-translation motion model, 227–230
Block-wise filtering, 158, 163–164
Blue-screen matting (Chroma keying), video capture, 329
Blur
   adaptive LMMSE filter avoiding, 155
   asymmetric coding and, 506–507
   cross-talk causing, 63
   down-conversion with anti-alias filtering and, 351
   image decimation and, 111
   image restoration undoing image, 169
   image/video quality and, 96
   SR in frequency domain and, 387, 389
   as tradeoff in LSI denoising filter, 150
Blur identification, and blind restoration, 175–176
Blur models
   overview of, 164
   point-spread function, 165–167
   shift-varying spatial blurring, 168
   space-invariant spatial blurring, 167
Blu-ray disc specification, 361
BM3D (block-matching 3D) filtering
   image de-noising with, 163–164
   image restoration with, 174
   V-BM3D extension for, 374
BM4D (block matching 4D) filtering, 374
Bob-and-weave filter, 357–358
Boundary
   in 2D recursive filtering, 28–29
   of changed regions in background subtraction, 295
   as image restoration problem, 175
   in JPEG2000, 451
   in phase-correlation method, 250
Box filtering, 106–108, 112
B-pictures
   in medium-grain SNR scalability, 501
   in MPEG-1, 469, 473–476
   in MPEG-2, 477–478, 482–483
   in temporal scalability, 498–499
Brightness
   in human vision, 58
   in pixel-based contrast enhancement, 137–138
Broken-link access (BLA) picture, HEVC, 492
B-slices, H.264/AVC, 485, 488
Bundle adjustment, projective reconstruction, 259–260
C
   CABAC (context-adaptive binary arithmetic coding), H.264/AVC, 489–490, 496
   Cable television
      analog video format in, 64
      ATSC standards for, 86
      DVB-C and DVB-C2 standards for, 88–89
   Calling mode, JPEG-LS, 427
   Camcorders, 462
   Camera calibration
      in 3D motion/structure estimation, 252–253
      in Dense SFS, 263
      matrix, 198–199, 260
   Camera projection matrices, 257
   Camera shake, and image blur, 167
   Camera-motion matrix, 254–255
   Canny edge detection, 134–135
<table>
<thead>
<tr>
<th>Term</th>
<th>Page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cartesian coordinates, 197–198, 200</td>
<td></td>
</tr>
<tr>
<td>CATV, DVB-C standard for, 89</td>
<td></td>
</tr>
<tr>
<td>CAVLC (context-adaptive variable-length coding), H.264/AVC, 489–490</td>
<td></td>
</tr>
<tr>
<td>CBs (coding blocks), 492–493</td>
<td></td>
</tr>
<tr>
<td>CCITT. See ITU-T (International Telecommunications Union)</td>
<td></td>
</tr>
<tr>
<td>CCMF (cross-correlated multi-frame) Wiener filter, 377</td>
<td></td>
</tr>
<tr>
<td>CCN (content centric networking), MPEG-DASH, 94</td>
<td></td>
</tr>
<tr>
<td>CDF (cumulative distribution function), 138, 141</td>
<td></td>
</tr>
<tr>
<td>CEA (Consumer Electronics Association), 76</td>
<td></td>
</tr>
<tr>
<td>Center of projection, 196–199</td>
<td></td>
</tr>
<tr>
<td>Central difference, estimating image gradient, 129</td>
<td></td>
</tr>
<tr>
<td>CFA (color filter array) interpolation, 120</td>
<td></td>
</tr>
<tr>
<td>Change detection, image segmentation</td>
<td></td>
</tr>
<tr>
<td>background subtraction, 291–298</td>
<td></td>
</tr>
<tr>
<td>overview of, 289</td>
<td></td>
</tr>
<tr>
<td>shot-boundary detection, 289–291</td>
<td></td>
</tr>
<tr>
<td>Change detection, in pel-recursive motion estimation, 246</td>
<td></td>
</tr>
<tr>
<td>Changing elements, in 2D RLC, 421–423</td>
<td></td>
</tr>
<tr>
<td>Checkerboard format, stereo-video, 83</td>
<td></td>
</tr>
<tr>
<td>Chroma keying (blue-screen matting), video capture, 329</td>
<td></td>
</tr>
<tr>
<td>Chrominance</td>
<td></td>
</tr>
<tr>
<td>composite-video format encoding, 66</td>
<td></td>
</tr>
<tr>
<td>perceived by human eye, 54–55</td>
<td></td>
</tr>
<tr>
<td>spatial-frequency response varying with, 60–61</td>
<td></td>
</tr>
<tr>
<td>S-video signal, 66</td>
<td></td>
</tr>
<tr>
<td>Chunks, HTTP, 93–94</td>
<td></td>
</tr>
<tr>
<td>CIE (International Commission on Illumination), 55, 161–162</td>
<td></td>
</tr>
<tr>
<td>CIERGB color space, 56</td>
<td></td>
</tr>
<tr>
<td>CIEXYZ (or sRGB) color space, 56, 57</td>
<td></td>
</tr>
<tr>
<td>Circle of confusion, out-of-focus blur, 165–166</td>
<td></td>
</tr>
<tr>
<td>Circular convolution, 26</td>
<td></td>
</tr>
<tr>
<td>Circular filters, image smoothing with LSI filter, 106</td>
<td></td>
</tr>
<tr>
<td>Circular shift, DFT, 17</td>
<td></td>
</tr>
<tr>
<td>Circular symmetry, 5, 27</td>
<td></td>
</tr>
<tr>
<td>Classic image/video segmentation, matting, 328</td>
<td></td>
</tr>
<tr>
<td>Clean random-access (CRA) picture, HEVC, 491–492</td>
<td></td>
</tr>
<tr>
<td>Closed GOPs, HEVC, 491–492</td>
<td></td>
</tr>
<tr>
<td>CLS (constrained least-squares) filtering, 170–173, 375–377</td>
<td></td>
</tr>
<tr>
<td>Clustering</td>
<td></td>
</tr>
<tr>
<td>adaptive MAP method of, 284–285</td>
<td></td>
</tr>
<tr>
<td>for image segmentation, 277–278</td>
<td></td>
</tr>
<tr>
<td>K-means algorithm for, 278–279</td>
<td></td>
</tr>
<tr>
<td>mean-shift algorithm for, 279–280</td>
<td></td>
</tr>
<tr>
<td>as multiple-motion segmentation, 302–306</td>
<td></td>
</tr>
<tr>
<td>CMYK (Cyan, Magenta, Yellow, Black) model, 56–57</td>
<td></td>
</tr>
<tr>
<td>Coarse-grain SNR scalability, 501</td>
<td></td>
</tr>
<tr>
<td>Coder, of source encoder, 405</td>
<td></td>
</tr>
<tr>
<td>Coding blocks (CBs), 492–493</td>
<td></td>
</tr>
<tr>
<td>Coding-tree units (CTUs), HEVC, 492–495</td>
<td></td>
</tr>
<tr>
<td>Coefficients, DCT, 434</td>
<td></td>
</tr>
<tr>
<td>Coiflet filters, 124–125</td>
<td></td>
</tr>
<tr>
<td>Collaborative filtering, BM3D, 164</td>
<td></td>
</tr>
<tr>
<td>Color</td>
<td></td>
</tr>
<tr>
<td>analog-video signal standards, 65–67</td>
<td></td>
</tr>
<tr>
<td>capture/display for digital video, 69–70</td>
<td></td>
</tr>
<tr>
<td>human visual system and, 54–56</td>
<td></td>
</tr>
<tr>
<td>image processing for digital video, 71–74</td>
<td></td>
</tr>
<tr>
<td>management, 57</td>
<td></td>
</tr>
<tr>
<td>Color balance, in image processing, 105</td>
<td></td>
</tr>
<tr>
<td>Color de-mosaicking, 120</td>
<td></td>
</tr>
<tr>
<td>Color filter array (CFA) interpolation, 120</td>
<td></td>
</tr>
<tr>
<td>Color space</td>
<td></td>
</tr>
<tr>
<td>calibrating devices to common, 57</td>
<td></td>
</tr>
<tr>
<td>color management and, 57</td>
<td></td>
</tr>
<tr>
<td>HSI (hue-saturation-intensity), 73–74</td>
<td></td>
</tr>
<tr>
<td>of JPEGs, 436–437</td>
<td></td>
</tr>
<tr>
<td>RGB and CMYK, 56–57</td>
<td></td>
</tr>
<tr>
<td>Color transforms, JPEG2000, 449</td>
<td></td>
</tr>
<tr>
<td>Color-matching functions, 54–56</td>
<td></td>
</tr>
<tr>
<td>Color-only pixel segmentation, 311–313</td>
<td></td>
</tr>
<tr>
<td>Color-region-based motion segmentation, 311–313</td>
<td></td>
</tr>
</tbody>
</table>
Cross-talk, auto-stereoscopic displays, 63, 81
CTUs (coding-tree units), HEVC, 492–495
Cubic-convolution interpolation filter, 116–118
Cumulative distribution function (CDF), 138, 141

Data partitioning, 454, 481, 490
Data rates
3D video, 79
digital video vs. digital audio, 71
multi-view video, 83
SDTV, 75
Data structure
H.264/AVC, 484–485
HEVC, 491–492
MPEG-1, 469–471
MPEG-2 video, 477–478
Daubechies (9,7) floating-point filterbank, JPEG2000, 450
dB (decibels), signal-to-noise ratio in, 149
DBBP (depth-based block partitioning), 3D-HEVC, 511
DC coefficient, 471–472
DC gain, 451
DC mode, 486, 495
DCI (Digital Cinema Initiative), 90
DCP (disparity-compensated prediction), 3D-HEVC, 511
DCPs (Digital Cinema Packages), 89
DCT (discrete cosine transform)
2D-DCT. See 2D-DCT (discrete cosine transform)
3D-DCT, 463–464
image/video compression artifacts, 97
medium-grain SNR scalability and, 501
MPEG-1 and, 471–475
MPEG-2 and, 479–481
overview of, 18–19
relationship to DFT, 19–20
video compression with, 462–463
DCT coding and JPEG
encoder control/compression artifacts, 442–443
JPEG baseline algorithm, 435–436
JPEG color, 436–437
JPEG image compression, 407–408
JPEG progressive mode, 441–442
JPEG psychovisual aspects, 437–441
JPEG standard, overview, 434–435
overview of, 431–434
Deadzone
JPEG2000, 451–452
mid-tread uniform quantizer, 407
De-blocking filter, 490, 497
Deblurring of images. See Image restoration
Decibels (dB), signal-to-noise ratio in, 149
Decimation, image, 111–113, 117
Decoded picture buffer (DPB), H.264/AVC, 485
Decoder-side view synthesis, 3D-HEVC, 512
Decoding
CABAC, 496
compression artifacts in, 442
H.264/AVC, 484–485, 490
HEVC parallel, 493–495
MPEG-1, 468–471, 476
MVC, 505–506
RGB sampled signal, 67
SHVC parallel, 498
Decomposition
into 1D transforms, 9–10
wavelet, 121–122, 126–127
Defocus blur, 96
Degradation from space-varying blurs, 177–180
Dehomogenization, perspective projection, 198
De-interlacing
critical velocities in, 354
inter-field temporal (weave filtering), 357–358
intra-field, 355–357
motion-compensated, 359–361
overview of, 46–47
in video-format conversion, 355
De-mosaicking, color, 120
Denoising, image
  image and noise models, 148–150
  local adaptive filtering, 153–157
  LSI filters in DFT domain, 150–153
  nonlinear filtering, bi-lateral filter, 161–162
  nonlinear filtering, median, 158–159
  nonlinear filtering, order-statistics filters, 159–160
  nonlinear filtering, wavelet shrinkage, 160–161
  non-local filtering, BM3D, 163–164
  non-local filtering, NL-Means, 162–163
  overview of, 147–148
De-noising video. See Multi-frame noise filtering
Dense SFS (structure from stereo), 263
Dense structure from zero, 263
Dense-correspondence estimation problem, 2D, 215–216
Dense-motion (optical flow/displacement) estimation, 215–216
Dense-motion estimation problem, 2D
  correspondence vectors in, 215–216
  optical flow vectors in, 216
  overview of, 215
Depth coding tools, 3D-AVC, 508–509
Depth map coding
  3D-HEVC, 510, 511
  MVC+D (MVC plus depth maps), 507
  in view-plus-depth format, 84–85
Depth perception, stereo vision, 62–63
Depth-based block partitioning (DBBP), 3D-HEVC, 511
Depth-based motion vector prediction (DMVP), 3D-AVC, 509
Depth-based representations, multi-view video, 83–84
Depth-image based rendering (DIBR), 84, 507, 509–510
Derived quantization mode, JPEG2000, 452
Deterministic non-parametric model, 213
Device-dependent color models, 56, 57
Device-independent color space, 57
DFD (displaced-frame difference)
  2D apparent-motion estimation, 219–220
  in Bayesian motion estimation, 247–249
  in MC filter post-processing, 349
  in motion-field model and MAP framework, 315
  in pel-recursive motion estimation, 246–247
DFT (discrete Fourier transform)
  in 2D DFT/inverse 2D DFT, 15–16
  and boundary problem in image restoration, 175
  convergence, 15
  DCT closely related to, 19–20, 433
  DCT preferred over, 18
  exercise, 51
  FIR filters and symmetry, 26
  image smoothing with LSI filter, 106
  implementing IIR filters, 28–29
  implementing LSI denoising filters, 150–153
  and JPEG. See JPEG standard
  normalized frequency variables, 15
  overview of, 14–15
  in phase-correlation motion estimation, 250
  properties of, 16–18
  pseudo-inverse filtering with, 169
Diamond search (DS), matching method, 237
DIBR (depth-image based rendering), 84, 507, 509–510
Dictionary size, in Lempel–Ziv coding, 431
Difference of Gaussian (DoG) filter, 134–135
Differential image, 408
Differential methods, motion estimation
  in deterministic non-parametric model, 213
  Horn–Shunck method, 230–233
  Lukas–Kanade method, 225–230
  overview of, 225
Differential pulse-code modulation. See DPCM (differential pulse-code modulation)
Diffusion-based-in-painting, image restoration, 180–181
Digital cinema, 89–92
Digital Cinema Initiative (DCI), 90
Digital Cinema Packages (DCPs), 89
Digital dodging-and-burning, tone mapping, 142–143
Digital images
defined, 1
finite quarter-plane support of, 3
Digital images and video
analog video, 63–67
definition of, 53
digital video. See Digital video
human vision and. See Human visual system/color
overview of, 53
quality factors, 96–100
Digital micromirror devices (DMDs), DLP projectors, 89
Digital Terrestrial Multimedia Broadcasting standards, 86
Digital video
3D video, 79–85
analog-to-digital conversion, 33–34, 66–67
applications, 85–95
color image processing, 71–74
color/dynamic range/bit depth in, 69–71
defined, 1
Digital TV (DTV) standards, 85–89
orthogonal sampling lattice for progressive, 34
revolution in, 67
spatial resolution/frame rate in, 67–69
standards, 74–78
vertically aligned 2:1 interlace lattice for, 34
Digital Video Broadcasting (DV) standards, 86
Digital Visual Interface (DVI) standard, 77
Digital-video applications
computer vision and scene/activity, 95
digital cinema, 89–92
digital TV (DTV), 85–89
video streaming over Internet, 92–95
Digitization, of analog, 64, 75
Direct Linear Transformation (DLT), 243–244
Direct segmentation, 307–309
Directional filtering, 20, 156–157
Directional smoothness constraints, 232–233
Discontinuity modeling, in Bayesian motion estimation, 248–249
Discrete cosine transform. See DCT (discrete cosine transform)
Discrete Fourier transform. See DFT (discrete Fourier transform)
Discrete memoryless source (DMS), 402–404
Discrete random process, 402
Discrete signals
definition of, 2, 6
discrete Fourier transform (DFT), 14–18
Fourier transform of, 12–14, 35
Discrete Weiner-Hopf equation, 152
Discrete-discrete model, low-resolution sampling, 384–386
Discrete-sine transform (DST), HEVC, 496
Discrete-trigonometric transform, DCT as, 433
Discrete-wavelet transform (DWT), 443, 448, 450
Disocclusion regions, 85
Disparity-compensated prediction (DCP), 3D-HEVC, 511
Displaced-frame difference. See DFD (displaced-frame difference)
Display order, MPEG-1, 470–471
Display technologies
3D, 79–82
classification of, 80
digital video standards, 76
Distortion, quantizer, 406
DLP projectors, 89, 90
dLT (Direct Linear Transformation), 243–244
DMDs (digital micromirror devices), DLP projectors, 89
DMS (discrete memoryless source), 402–404
DMVP (depth-based motion vector prediction), 3D-AVC, 509
Dodging-and-burning, image enhancement, 142–143
DoG (Difference of Gaussian) filter, 134–135
Dolby 3D cinema technology, 91
Domain filtering, 161–162, 164
Dominant-motion segmentation, 296, 299–302
Double stimulus comparison scale (DSCS), 98
Double stimulus continuous quality scale (DSCQS), 98
Double stimulus impairment scale (DSIS), 98
Down-conversion
with anti-alias filtering, 351–352
sampling structure conversion, 43–45
in video-format conversion, 351
without anti-alias filtering, 352–353
Down-sampling (sub-sampling)
in down-conversion, 351–354
of frame-compatible formats, 83, 503
in image decimation, 111–113
sampling structure conversion, 44
DPB (decoded picture buffer), H.264/AVC, 485
DPCM (differential pulse-code modulation) hybrid DPCM video coding, 464
JPEG baseline mode, 436, 439
MC-DPCM video compression, 466
D-pictures, in MPEG-1, 469
Drift
avoiding in particle-filter motion tracking, 325
in template-tracking methods, 321
DS (diamond search), matching method, 237
DSCQS (double stimulus continuous quality scale), 98
DSCS (double stimulus comparison scale), 98
DSIS (double stimulus impairment scale), 98
DSM-CC, MPEG-2, 476
DSS (Hughes), satellite transmission, 87
DST (discrete-sine transform), HEVC, 496
DTV (Digital TV) standards, 85–89
Dual-prime prediction for P-pictures, MPEG-2, 479
DV (Digital Video Broadcasting) standards, 86
DVB standards, 87–89
DVB-H standard, 89
DVB-S, satellite transmission, 87
DVB-S2X, 88
DVCAM format, 462
DVCPRO/DVCPRO-50/DVCPRO HD formats, 462
DVI (Digital Visual Interface) standard, 77
DWT (discrete-wavelet transform), 443, 448, 450
Dyadic structure, temporal scalability, 498–499
Dynamic range
compression, 143
expansion, 137
overview of, 70

E
Early lossless predictive coding, 424–426
EBCOT (embedded block coding with optimized truncation), 448
EBU (European Broadcasting Union), DVB, 87–89
Edge detection
Canny, 134–135
Harris corner detection, 135–137
operators, 130–131
overview of, 127–128
Edge preserving filtering
with adaptive LMMSE filter, 155–156
bi-lateral filters, 162
with directional filtering, 156–157
with median filtering, 158–159
Edge-adaptive filters, 20, 118–119
Edge-adaptive intra-field interpolation, 356–357
Edges, modeling image, 127
Eigenvalues, Harris corner detection, 136
EM (expectation-maximization) algorithm, 174, 294
Embedded block coding with optimized truncation (EBCOT), 448
Embedded quantization, JPEG2000, 452
Embedded zerotree wavelet transform (EZW), 448
Encoder control, 442–443, 511–512
Encoding
HEVC, 493–495
HEVC parallel, 493–495
in MPEG-1, 470–471, 475–476
in MPEG-2 video, 482–483
Encrypted DCP files, 89
Endpoint error, motion estimation, 223–224
Energy-compaction property, DCT, 433–434
Enhanced-depth coding tools, 3D-AVC, 509
Enhancement, image
   classifying methods of, 137
   with pixel-based contrast, 137–142
   with spatial filtering, 142–147
Entropy coding
   H.264/AVC improvements to, 489–490
   HEVC, 496
   Huffman coding as. See Wavelet representations
   JPEG 2000, 453
   of symbols, 410
Entropy of source, and lossless encoding, 403
Entropy-constrained quantizers, 407
Epipolar geometry, 255–257
Equalization, histograms, 140–142
Error measures, motion estimation, 223–224
Error resilience
   H.264/AVC tools for, 490–491
   JPEG 2000, 454
ETSI (European Telecommunications Standards Institute), 87–89
Euclidean reconstruction, 260
Euclidean structure, 252–254
Euler-Lagrange equations, motion estimation, 230–231
European Broadcasting Union (EBU), DVB, 87–89
Exemplar-based methods, image-in-painting, 181
Expectation-maximization (EM) algorithm, 174, 294
Experts, subjective quality assessments, 97–98
Expounded quantization mode, JPEG2000, 452
Extended Graphics Array (XGA), 76
Extended profiles, H.264/AVC, 484
Extrinsic matrix, perspective projection, 198
Eyes. See Human visual system/color
EZW (embedded zerotree wavelet transform), 448

F
Fast Fourier Transform (FFT) algorithms, 15–16, 18–20
Fast search, in block-matching, 236–238
Fax standards, digital, 78
Fax transmission, with RLC, 419
FD (frame difference), background subtraction, 291–293
Feature correspondences, in homography estimation, 242–245
Feature-tracking methods, 318–321
Felzenszwalb–Huttenlocher method, graph-based video segmentation, 319
FFT (fast Fourier Transform) algorithms, 15–16, 18–20
Fidelity range extensions (FRExt), H.264/AVC, 483–484, 486, 489
Field pictures, MPEG-2, 477–480
Field prediction for field/frame pictures, MPEG-2, 479
Field rate, video signal, 65, 349
Field sequential format, stereo-video compression, 503–504
Field-DCT option in MPEG-2, 479–481, 483
Film-mode detection, frame-rate conversion, 367
Filterbanks, JPEG2000, 450
Filtered-model frame, 292
Filter-frequency response, image smoothing, 106
Filters
   adaptive, 113, 145–146, 155–157
   FIR. See FIR (finite-impulse response) filters
   H.264/AVC in-loop de-blocking, 490
   JPEG2000 normalization, 451
   JPEG2000 wavelet, 450
   multi-frame noise. See Multi-frame noise filtering
   video. See Video filtering
   wavelet transform coding and, 443–447
   zero-order hold, 115–116, 360–361
Finite differences, noise sensitivity of, 129
Finite extent signals, 2–3, 5, 14
Finite-difference operators, 128–131, 133–134
Finite-support signal, 6
Index

FIR (finite-impulse response) filters
designing cubic-convolution filter with, 117
JPEG2000, 450
in LSI filtering for constant-velocity global motion, 347
LSI systems, 20
and symmetry in MD systems, 25–27
wavelet representations. See Wavelet representations
wavelet transform coding, 445
for zero or linear phase in image processing, 24
Firewalls, HTT streaming and, 93
FIR-LLMSE filter, 150–151
FIR-LMMSE filter, 154–155
First derivatives, edge of 1D continuous signal, 127–128
Fixed-length coding, symbol encoding, 409–410
Fixed-reference frame, in background subtraction, 292
Flexible macroblock ordering (FMO), H.264/AVC, 490
Flicker frequency, and human eye, 61
Floating-point filterbanks, JPEG2000, 450
FMO (flexible macroblock ordering), H.264/AVC, 490
Formats
analog-video signal, 65–66
conversion of video. See Video-format conversion
digital-video spatial resolution, 68–69
stereoscopic video, 82–83
Forward prediction, MPEG-1, 472
Forward-motion estimation, 215–216
Four-fold symmetry, 5, 14, 20
Fourier transform
of continuous signals, 8–12
discrete. See DFT (discrete Fourier transform)
of discrete signals, 12–14, 35
exercises, 50
frequency response of system and, 24
impulse response of low-pass filter, 27
Nyquist criterion for sampling on lattice, 37
of signal sampled on lattice, 34–36
SR in frequency domain with, 386–387
Four-step logarithmic search, matching method, 237
Fovea, of human eye, 54, 62
FR (full reference metrics), 98–100
Frame difference (FD), background subtraction, 291–293
Frame packing, HDMI, 83
Frame pictures, MPEG-2 video, 477–478
Frame prediction for frame pictures, MPEG-2 video, 479
Frame rate
digital video, 67–69, 90
measuring temporal-frequency response, 61
temporal scalability and, 498–499
video signal, 65
video-format conversion. See Frame-rate conversion, video-format
Frame sequential format, stereo-video compression, 503–504
Frame-compatible formats, stereoscopic video, 82–83, 503–504
Frame-DCT, MPEG-2, 479–481, 483
Frame-rate conversion, video-format
24 Hz movies to 50/60 Hz, 361–363
50 to 60 Hz, 363
definition of, 361
film-mode detection, 367
MC frame/scan-rate conversion, 366
motion-adaptive scan-rate conversion, 365–366
scan-rate doubling, 363–365
Frames, motion segmentation using two, 299–300
Free-view 2D video, 503
Freeze frame, 97
Frequency domain
analyzing LSI denoising filters in, 150
sampling on MD lattices, 36–41
super-resolution in, 386–389
unsharp masking image enhancement and, 144
Frequency response
1D binary decomposition filters, 122
convolution in Fourier domain, 25
IIR Weiner filter, 152–153
MC filter, 346
MD systems, 23–25
out-of-focus blur, 166
sampling structure conversion, 44–45
of zero-order-hold filter, 115
Frequency shifting property, DCT, 19
Frequency shifting property, DFT, 17
Frequency spectrum of video, 342–345
Frequency variables, 8, 12, 14–15
FRExt (fidelity range extensions), H.264/AVC,
483–484, 486, 489
Full reference metrics (FR), 98
Full search, in block-matching, 235–236, 241
Full-resolution stereo-video format, 83
Fundamental matrix, projective reconstruction,
255–257

G
Games, stereo-video format for, 83
Gaussian filters
anisotropic, 119
bi-lateral filters and, 162
Canny edge detection and, 134
directional filtering using, 157
estimating Laplacian of, 134–135
estimating partials by derivatives of, 131–132
extending with bi-lateral filtering, 109–110
in image decimation, 112
image smoothing with LSI filter and, 106,
108–109
Gaussian noise, 148
Gaussian pyramid
in hierarchical motion estimation, 223–224,
240
JPEG hierarchical, 442
in multi-resolution frame difference analysis,
293
overview of, 120
Gaussians, in background modeling, 293–295
Gaze-shifting eye movements, 62
General model (depth map), 208
General motion-compensated de-interlacing,
360–361
General periodic signal, in 2D, 4
Generalized block-matching, motion estimation,
241–242
Generalized convergence, 9, 12
Geometric image formation, 196–199, 202
Ghosting, from cross-talk, 63
Gibbs potential, 315
Gibbs random field (GRF), 285–286
Glasses, stereoscopic 3D, 91
Global positioning system (GPS) motion
tracking, 317
Global thresholds, 276
Global translation, constant-velocity, 343–345
Global-MC de-interlacing, 357–358
Golomb–Rice coding, 428–429
GOP (group of pictures)
decoding in MVC, 506
HEVC, 491–492
in medium-grain SNR scalability, 501
in MPEG-1, 469–471
Gradient estimation/edge/features
Canny edge detection, 134–135
Harris corner detection, 135–137
of image gradient, 128–131
of Laplacian, 132–134
overview of, 127–128
of partials, 131–132
Gradual view refresh (GVR), 3D-AVC/MVC+D,
509
Graph-based methods
image segmentation, 285–287
spatio-temporal segmentation/tracking, 319
Gray codes, symbol encoding, 409–410
GRF (Gibbs random field), 285–286
Ground-truth data. See GT (ground-truth) data
Group of pictures. See GOP (group of pictures)
Groups, BM3D filtering, 164
GT (ground-truth) data
3D coordinates, Euclidean reconstruction, 260
motion vectors, motion estimation, 224–225
segmentation/tracking, 330–331
GVR (gradual view refresh), 3D-AVC/MVC+D, 509

H
H.261, 467–468
H.262, 467
H.263, 467
H.264, 88
H.264/AVC (MPEG-4 AVC/ITU-T H.264) standard
  3D-video compression overview, 503
  input-video format/data structure, 484–485
  intra-prediction, 485–486
  motion compensation, 486–488
  MVD compression and, 507
  other tools and improvements, 489–491
  overview of, 483–484
  stereo/multi-view video-coding extensions of, 504–507
  stereo-video SEI messages in, 504
  temporal scalability, 498
  transform, 488–489
H.265/HEVC standard, 504, 507
Half-plane support, 2–3, 8
Harris corner detection, 135–137
HCF (highest confidence first) algorithm, 316–317
HDMI (High-Definition Multimedia Interface), 77, 83
HDR (high dynamic range), 70
HDS (HTTP Dynamic Streaming), Adobe, 94
HDTV (high-definition TV), 75–78, 88
Head-motion parallax, 81
Head-motion range, 81
Hermitian symmetric, 13–14, 16–17
Hessian matrix, 126
HEVC (high-efficiency video-coding) standard
  3D-HEVC tools, 510–512
  3D-video compression overview, 503
coding-tree units, 492–493
entropy coding, 496
intra-prediction, 495
in-loop de-blocking filter, 497
motion compensation, 495–496
motion vector coding, 496
overview of, 491
parallel encoding/decoding tools, 493–495
transform and quantization, 496
video-input format and data structure, 491–492
Hexagonal matching, 241
Hexagonal sampling, 33
HFR (high frame rate), digital video, 90
Hi10P (High 10 Profile), H.264/AVC, 484
Hi422P (High 4:2:2 Profile), H.264/AVC, 484
Hi444P (High 4:4:4 Profile), H.264/AVC, 484
Hierarchical Bayesian motion-estimation, 248
Hierarchical block matching, 234, 240–241
Hierarchical iterative-refinement, Lukas–Kanade motion estimation, 226–230
Hierarchical mode JPEG, 442
Hierarchical motion estimation, 223–224
Hierarchical prediction structures, 498–499
High 4:2:2 Profile (Hi422P), H.264/AVC, 484
High 4:4:4 Profile (Hi444P), H.264/AVC, 484
High 10 Profile (Hi10P), H.264/AVC, 484
High dynamic range (HDR), 70
High frame rate (HFR), digital video, 90
High Profile (HP), H.264/AVC, 484
High profiles
  H.264/AVC, 484, 486
  MPEG-2, 482
High-Definition Multimedia Interface (HDMI), 77, 83
High-definition TV (HDTV), 75–78, 88
High-efficiency video-coding. See HEVC (high-efficiency video-coding) standard
Highest confidence first (HCF) algorithm, 316–317
High-level computer vision, 95
High-pass filter, wavelet analysis, 122–123
High-resolution (HR) image, super-resolution, 378
Histogram
change detection methods, 290
normalization (contrast stretching), 137
pixel-based contrast enhancement, 138, 140–142
HLS (HTTP Live Streaming), Apple, 94
Holographic displays, 82
Homography (perspective model), 208–209, 263
Homography estimation, of motion, 229–230, 243–245
Homomorphic filtering, image enhancement, 146–147
Horizontal mode, two-dimensional RLC, 421–423
Horizontal spatial-frequency pattern, MD signals, 6
Horizontal sum buffer (HSB), box filtering, 107–108
Horn–Shunck motion estimation, 230–233, 249
Hough transform method, 305–306
HP (High Profile), H.264/AVC, 484
HP (hi-pass) temporal bands, 465
HR (high-resolution) image, super-resolution, 378
HSB (horizontal sum buffer), box filtering, 107–108
HSI (hue-saturation-intensity), color image, 73–74
HTTP (hyper-text transport protocol), streaming over, 93–94
HTTP Dynamic Streaming (HDS), Adobe, 94
HTTP Live Streaming (HLS), Apple, 94
Huber-Markov random field model, 119
Hue-saturation-intensity (HSI), color image, 73–74
Huffman coding
block, 414
in early lossless predictive coding, 424
as entropy coding, 410
Golomb–Rice coding equivalent to, 429
in image compression, 410–413
lossless coding theorem and, 404
one-dimensional RLC and, 419–421
Human visual system/color
asymmetric coding of stereo video and, 506–507
color vision and models, 54–57
computer vision aiming to duplicate, 95
contrast sensitivity, 57–59
overview of, 54
spatio-temporal frequency response, 59–62
stereo/depth perception, 62–63
Hybrid DPCM video coding, 464
Hybrid methods, MC de-interlacing, 361
Hybrid scalability, SVC compression, 502
Hyper-text transport protocol (HTTP), streaming over, 93–94
Hysteresis thresholding, 135

I
IC (illumination compensation), 3D-HEVC, 511
ICC (International Color Consortium) profiles, 56, 57
ICM (iterated conditional mode), 283–285, 316–317
ICT (irreversible color transform), JPEG2000, 449
IDCT (inverse DCT), MPEG-1, 475
IDD-BM3D (iterative decouple deblurring-BM3D), 174
Identity operator, regularization by CLS, 173
IDR (instantaneous decoding refresh) picture, H.264/AVC, 491
IETF (Internet Engineering Task Force), streaming protocols, 93
IIR (infinite-impulse response) filters, 20–21, 27–29
IIR Weiner filter, 151–153
IIR-LLMSE filter, 151–153
Ill-posed problem, 220–223
Illumination
homomorphic filtering and, 146–147
retinex for removing undesirable, 146
Illumination compensation coding tool, MVC, 506
Illumination compensation (IC), 3D-HEVC, 511
Image and video matting, 328–329
Image capture, 96
Image compression
  arithmetic coding, 414–417
  DCT, 432–435
  definition of, 401
  and digital video, 71
  elements of image compression systems, 405–406
  exercises, 456–459
  Huffman coding, 410–414
  information theoretic concepts, 402–405
  ISO JPEG2000 standard, 448–454
  lossless. See Lossless image compression methods, 405–406
  orthogonality in, 125
  overview of, 401–402
  quantization, 406–409
  symbol coding, 409–410
  wavelet transform coding, 443–448
Image filtering
  denoising. See Denoising, image
  enhancement. See Enhancement, image
  exercises, 186–193
  gradient estimation/edges/features. See
    Gradient estimation/edge/features
  overview of, 105
  re-sampling/multi-resolution. See Re-sampling restoration. See Restoration, image
  smoothing, 106–110
Image formation
  with affine camera, 199–201
  overview of, 196
  photometric effects of 3D motion, 201–202
  with projective camera, 196–199
Image gradients, 176
Image in-painting, restoration, 180–181
Image matting, 329
Image models, and performance limits, 149–150
Image noise, 96
Image processing, 24, 105
Image quality. See Video/image quality
Image rectification, in Dense SFS, 263
Image registration problem, 217
Image re-sampling. See Re-sampling
Image restoration. See Restoration, image
Image segmentation
  active contour models (snakes), 287–289
  Bayesian methods, 277–285
  clustering, 277–280
  graph-based methods, 285–287
  overview of, 275
  thresholding, 275–277
Image sharpening. See Spatial filtering
Image smoothing. See Smoothing, image
Image- and video-compression standards, digital video, 77–78
Imaginary part, frequency response, 24
IMAX Digital 3D cinema technology, 91–92
Impulse, MD, 6–7, 49
Impulse response
  in 2D convolution, 20–23
  in 2D recursive filtering, 29
  of circularly symmetric low-pass filter, 27
  of cubic-convolution filter, 117
  of ideal interpolation filter, 115
  of linear interpolation filter, 116
  of MC filter, 346
  in polyphase of interpolation, 117
  of zero-order-hold filter, 115
Inertial sensing motion tracking, 317
Infinite-impulse response (IIR) filters, 20–21, 27–29
Information theoretic concepts, 402–405
Initialization, 296, 324
In-Loop Block-Based View Synthesis Prediction (VSP), 3D-AVC, 509
In-loop de-blocking filter, 490, 497
Input signal, 111–114
Input-video format
  H.264/AVC, 484–485
  MPEG-1, 469–471
  MPEG-2 video, 477–478
Instantaneous decoding refresh (IDR) picture, H.264/AVC, 491
Integer transforms, H.264/AVC, 489
Integer-valued vectors, periodic signals, 4
Integrated Multimedia Broadcasting (ISDB) standards, 86
Intensity, HSI and, 73–74
Interactive semi-automatic segmentation, 329
Inter-coding, in MPEG-2, 483
Inter-field line averaging, scan-rate conversion, 364–365
Inter-field temporal (weave filtering) deinterlacing, 357–358
Inter-frame (temporal) redundancies, 461, 462
Inter-frame compression modes, MPEG-1, 472–474
Interlace lattice, 34
Interlace video input, 345
Interlaced scanning, analog-video signals, 64–65
Interlaced video, 469, 477–480
Inter-layer spatial scalable coding, 499–500
Interleaved format, stereo-video, 83
Interleaved ordering, JPEGs, 437
Intermediate signal, in image decimation, 111–112
International Color Consortium (ICC) profiles, 56, 57
International Commission on Illumination (CIE), 55, 161–162
International Telecommunications Union. See ITU-T (International Telecommunications Union)
International video compression standards, 467
Internet Engineering Task Force (IETF), streaming protocols, 93
Inter-object disparity measures, segmentation/tracking, 330
Interoperability, with video-format conversion, 349
Interpolation adaptive/nonlinear, 118–119
color de-mosaicking in, 120
cubic-convolution, 116–117
efficient polyphase implementation of, 117
interlacing and, 46–47
linear, 116
overview of, 113–115
sampling rate change by rational factor in, 117–118
single-frame SR in, 119
super-resolution vs. image, 379
zero-order-hold filter in, 115–116
Interpupilar distance, of average human, 62
Intersection of two lattices, 43
Inter-view coding with ALC, 3D-AVC, 509
Inter-view motion prediction, 3D-HEVC, 511
Intra-coding, MPEG-2, 483
Intra-field de-interlacing, video-format conversion, 355–357
Intra-field line averaging, scan-rate conversion, 364–365
Intra-frame compression modes, MPEG-1, 471–472
Intra-frame image-restoration problem, 168
Intra-frame video compression, 462–463
Intra-layer spatial scalable coding, 499–500
Intra-prediction
  H.264/AVC, 485–486
  HEVC, 495
Inverse 2D DFT, 15–16
Inverse 2D Fourier Transform, 8, 12
Inverse 2D Fourier transform, 34–35
Inverse DCT (IDCT), MPEG-1, 475
Inverse filtering, 169–170
Inverse Fourier transform, 35
Inverse pull-down, frame-rate conversion, 362–363
Inverse quantization at decoder, JPEG2000, 452–453
Inverse wavelet transform, in image denoising, 160
IP multicast, 95
I-pictures in MPEG-1, 469, 475
in MPEG-2 video, 477–479
Irregular Repeat-Accumulate codes, DVB-S2, 88
Irreversible color transform (ICT), JPEG2000, 449
Index

I-slice, H.264/AVC, 484
ISO (International Standards Organization)
  JPEG. See JPEG standard
  JPEG2000. See JPEG2000 standard
  video-compression standards, 467
Isometry, affine model, 212
Isomorphic signals, 4
Iterated conditional mode (ICM), 283–285, 316–317
Iterative decouple deblurring-BM3D (IDD-BM3D), 174
Iterative methods, in space-varying restorations, 177
ITU-T (International Telecommunications Union)
  G3 and G, 419–423
  H.261 standard, 467
  H.264/AVC. See MPEG-4 AVC/ITU-T H.264 standard
  ITU-R broadcast standards, 74–77
  sampling standards, 66–67
  standardizing digital-image communication, 78
  Video Quality Experts Group, 100
  video-compression standards of, 467

J
JBIG (Joint Bi-level Image Experts Group), 423–424
Johnson filters, 125
Joint Video Team (JVT). See MPEG-4 AVC/ITU-T H.264 standard
JP2 file format, JPEG2000, 454
JPEG standard
  baseline algorithm, 435–436
  color, 436–437
  DCT coding and, 431–434
  DV compression algorithm vs., 462–463
  encoder control/compression artifacts, 442–443
  as first still-image compression method, 78
  hierarchical mode, 442
  integer prediction in lossless mode, 425–426
  lossless mode of original, 424–426
  overview of, 434–435
  progressive mode, 441–442
  psychovisual aspects, 437–441
  uniform quantization in compression, 407–408
JPEG2000 standard
  boundary extension, 451
  for digital cinema, 78
  entropy coding, 453
  error resilience, 454
  filter normalization, 451
  inverse quantization at decoder, 452–453
  modern wavelet compression methods influencing, 448
  Motion JPEG 2000, 463
  overview of, 448
  prep-processing and color transforms, 449
  quantization at encoder, 451–452
  region-of-interest coding, 454
  wavelet filters, 450
JPEG-LS standard
  determination of calling mode, 427
  lossless image compression with, 426
  predictive-coding mode, 427–429
  run-length coding mode, 429–430
JVT (Joint Video Team). See MPEG-4 AVC/ITU-T H.264 standard

K
Kernel selection, directional filtering, 157
Key pictures
  H.264/AVC, 485
  medium-grain SNR scalability, 501
KLT (Kanade–Lucas–Tomasi) tracker, 319–321
KLT (Karhunen–Loeve transformation), 432
K-means algorithm, 278–280, 284, 302–305
K-nearest neighbor method, 280, 374

L
L (long) wavelength, and retinal cone, 54
Label assignment, clustering, 304–305
Labels, segmentation, 302–307
Lambertian reflectance model, 3D motion, 201–202

Laplacian
  determining edge pixels by, 134
  estimating by finite-difference operators, 132–133
  of Gaussian filter, 134–135
  in graph-based image segmentation, 287
  JPEG2000 quantization at encoder, 451–452
  in modeling image edges, 127
  regularization by CLS, 173

Laplacian of Gaussian (LoG) filter, 134–135

Last-first property, Lempel–Ziv codes, 430–431

Lattice(s), MD
  defining intersection of two, 43
  defining sum of two, 43
  definition of, 30
  Nyquist criterion for sampling on, 36–41
  reconstruction from samples on, 41–42
  sampling on, 30–34
  spectrum of signals sampled on, 34–36

LCD (liquid crystal) shutters, stereoscopic 3D glasses, 91

LCD monitors, flicker and, 76

Least significant bit-plane (LSB), bit-plane coding, 418

Least-squares (LS) solution, pseudo-inverse filtering, 170

Lempel–Ziv coding, lossless image compression, 430–431

Lenticular sheets, auto-stereoscopic multi-view displays, 80–81

Levels
  HEVC, 497
  MPEG-2 video, 477, 482

Lexicographic order, IIR filters, 28

Light, color models and, 56
  Light field displays, 80–82
  Line averaging, 356, 364
  Line repetition, intra-field de-interlacing, 355–356

Linear, shift-invariant filters. See LSI (linear, shift-invariant) filters

Linear approximations, to perspective-motion model, 212–213

Linear contrast manipulation, pixel-based enhancement, 139–140

Linear forward wavelet transform, image denoising, 160

Linear interpolation filter, 116, 355–356, 379

Linear minimum mean-square error (LMMSE) filter. See LMMSE (linear minimum mean-square error) filter

Linear phase, in image processing, 24–27

Linear shift-varying blurs. See LSV (linear shift-varying) blurs

Linear-motion blur, 166–167, 374–375

Liquid crystal (LCD) shutters, stereoscopic 3D glasses, 91

Live Internet broadcasting (push application), 92

Lloyd–Max quantizer, 406, 408

LMMSE (linear minimum mean-square error) filter
  adaptive, 155–157
  directional filtering, 157
  IIR Weiner filter, 151–153
  as optimal LSI denoising filter, 150–151
  video de-noising with adaptive, 370–372
  video de-noising with AWA filter vs., 373
  Wiener filter giving, 173

Local adaptive filtering
  adaptive LMMSE filter, 155–157
  FIR-LMMSE filter, 154–155
  image denoising with, 153

Local deformable motion, block-matching, 241–242

Local region-growing, image-in-painting, 181

Local thresholds, 276

LoG (Laplacian of Gaussian) filter, 134–135

Logarithmic search, matching method, 236–237

Log-likelihood, maximum-likelihood segmentation, 306–309

Log-luminance domain, retinex, 146

Long (L) wavelength, and retinal cone, 146
Lossless image compression
  adaptive arithmetic coding and JBIG, 423–424
  bit-plane coding, 418–419
  coding theorem, 403–404
  defined, 405–406
  early lossless predictive coding, 424–426
  JPEG-LS standard, 426–430
  Lempel–Ziv coding, 430–431
  overview of, 417
  RLC and ITU G3/G4 standards, 419–423
Lossy compression methods
  DCT and JPEG. See DCT coding and JPEG
  JBIG2 enabling for binary images, 423
  lossless vs., 405
  minimizing bit-rate, 404
  quantization as. See Quantization
  reversible color transform as, 449
  transform coding as basis of standards for,
      402, 431
  Low-delay structure, temporal scalability,
      498–499
  Low-level computer vision, 95
  Low-pass (LP) temporal bands, 3D-wavelet/sub-band coding, 465
  Low-pass filters
    image enhancement with unsharp masking,
        144–145
    KSI image smoothing, 106–109
    LSI denoising filter, 150
    LSI interpolation, 113–115
    prior to down-conversion, 47
    reconstruction from samples on lattice, 41–42
    sampling structure conversion, 44
    in wavelet analysis, 122–123
  Low-resolution (LR) frames, 378–380, 381–386
  LP (low-pass) temporal bands, 3D-wavelet/sub-band coding, 465
  LR (low-resolution) frames, 378–380, 381–386
  LS (least-squares) solution, pseudo-inverse
    filtering, 170
  LSB (least significant bit-plane), bit-plane
coding, 418
LSI (linear, shift-invariant) filters
  in constant-velocity global motion, 346–347
  convolution in Fourier domain, 24–25
  denoising in DFT domain, 150–153
  frequency response of system, 23–25
  image smoothing with, 106–109
  impulse response and 2D convolution in MD,
      21–22
  interpolation process with, 113–115
  up-conversion of video with MC, 352–353
LSV (linear shift-varying) blurs
  boundary problem in, 175
  image restoration problem of, 168
  overview of, 168, 169
  POCS framework, 177
  problem of, 168
  pseudo-inverse filtering, 169–170
  regularization by sparse-image modeling,
      173–174
  regularized deconvolution methods,
      170–173
  space-varying blurs, 177–180
  transforming degradations into LSI
degradations, 177
Lukas–Kanade motion estimation, 225–230
Luminance
  contrast sensitivity in human vision, 58
  in dodging-and-burning, 143
  dynamic range and, 70
  in MPEG-2 video, 477
  perceived by human eye, 54–55
  spatial-frequency response varying with,
      60–61
  spatio-temporal frequency eye response and,
      61–62
  S-video signal, 66
Luminance-chrominance
  color model, 71–72
  demonstration of JPEG-baseline algorithm,
      440–441
  JPEG psychovisual aspects and, 437–438
  JPEG supporting, 435, 436
Luminous efficiency function, CIE, 55
M
M (medium) wavelength, retinal cone sensitivity to, 54
Mach band effect, human vision, 58
Machine-learning methods, single-frame SR, 119
Macroblocks. See MBs (macroblocks)
MAD (minimum mean absolute difference), 234–235, 241
Magnitude, frequency response, 24
Main profiles
H.264/AVC, 484
HEVC, 497
MPEG-2, 482
MAP (maximum a posteriori) probability estimates
in adaptive/nonlinear interpolation, 119
Bayesian motion estimation, 247–249
Bayesian segmentation methods, 281–285
in blur identification from image gradients, 176
image restoration methods, 169
in POCS framework, 177–178
regularized deconvolution methods, 170–171
restoring images from LSI blur, 171
simultaneous motion estimation/segmentation, 314–315
for super-resolution reconstruction problem, 391
in Wiener filter, 173
Mapped error value (MErrval). Golomb-Rice coding, 429
Markov random field (MRF), adaptive image processing, 153
Marr-Hildreth scale space theory, 108
Masking, visual, 58–59
Matching methods, motion estimation
block-matching method, 234–238
generalized block-matching, 241–242
hierarchical block matching, 240–241
homography estimation, 243–245
overview of, 233–234
variable-size block matching, 238–240
Matting, image/video, 328–329
Maximum a posteriori. See MAP (maximum a posteriori) probability estimates
Maximum matching pel count (MPC), block matching, 234
Maximum transfer unit (MTU), 484–485
Maximum-displacement estimate, 250
Maximum-likelihood segmentation, 306–309
Maxshift method, ROI coding, 454
m-bit Gray code, bit-plane coding, 418
MBs (macroblocks)
H.264/AVC, 484–486
HEVC coding-tree units replacing, 492–493
in MPEG-1, 469–470
in MPEG-1 encoding, 475–476
in MPEG-1 inter-frame compression, 472–474
in MPEG-1 intra-frame compression, 471–472
in MPEG-2 video, 477–482
in MVC coding tools, 506
MC (motion compensation)
with connectivity constraints, 242
H.264/AVC, 486–488
HEVC, 495–496
MC-DPCM, 466
MPEG-1, 468–469, 472–476
overview of, 466
without connectivity constraints, 241–242
MC (motion-compensated) de-interlacing, 359–361
MC (motion-compensated) filtering
arbitrary-motion trajectories and, 345–346
AWA filter, 372–373
errors in motion estimation, 347–348
fpr multi-frame noise, 369–372
frame/scan-rate conversion, 366
general MC de-interlacing, 360–361
global-MC de-interlacing, 359
LSI filtering, 346–347, 352–353
MC-LMMSE filter, 370–372
motion estimates/occlusion handling, 349
overview of, 345
reliable motion estimation, 348
  in video processing, 20
MC (motion-compensated) interpolation, 47
MC zero-order hold filtering, 360–361
McCann99 retinex, 172
MCE (motion-compensation error), 223–224
MCMF (motion-compensated multi-frame) filter, 377
MCP (motion-compensated prediction)
  H.264/AVC, 485–488
  in MPEG-1, 468, 472, 474
  in MPEG-2 interlaced-video compression, 479
MC-transform coding, 466, 467
MD (multi-dimensional) sampling theory
  Nyquist criterion for sampling on lattice, 36–41
  overview of, 30
  reconstruction from sampling on lattice, 41–42
  sampling on lattice, 30–34
  spectrum of signals sampled on lattice, 34–36
  structure conversion, 42–47
MD (multi-dimensional) signals, 1–2, 5–8, 48
MD (multi-dimensional) systems
  FIR filters and symmetry, 25–27
  frequency response, 23–25
  IIR filters and partial difference equations, 27–29
  impulse response and 2D convolution, 20–23
  overview of, 20
MD (multi-dimensional) transforms
  Discrete Cosine Transform (DCT), 18–20
  Discrete Fourier Transform (DFT), 14–18
  Fourier transform of continuous signals, 8–12
  Fourier transform of discrete signals, 12–14
  overview of, 8
Mean filters, median filter vs., 159
Mean square difference (MSE), 99, 234, 241
Mean-shift (MS) algorithm, 279–280, 321–322
Mean-shift motion tracking, 321–322
Mean-square convergence, 9, 12
Mean-square quantization errors, 406, 408
Measurement matrix, multi-view Affine reconstruction, 254–255
Median filtering
  adaptive, 160
  denoising using, 158–159
  as energy function, 233
  in motion-adaptive scan-rate conversion, 365–366
  weighted, 159
Medical imaging modalities, projection slice theorem, 12
Medium (M) wavelength, retinal cone sensitivity to, 54
Medium-grain SNR (MGS) scalability, 501–502
Memory-management control-operation (MMCO), H.264/AVC, 488
MErrval (mapped error value). Golomb-Rice coding, 429
MGS (medium-grain SNR) scalability, 501–502
Mid-tread uniform quantizer, 407
Minimal-cut criterion, graph-based image segmentation, 286–287
Minimax, 161
Minimum mean absolute difference (MAD), 234–235, 241
Mixed MD signal, 2
MJ2K (Motion JPEG 2000), 463
ML (maximum likelihood) estimate, 159, 259–260
MMCO (memory-management control-operation), H.264/AVC, 488
Modulation schemes, digital video, 87–89
Monte Carlo method, 283, 323–325
Mosaic representation (image stitching), 217
Mosquito noise, 96
Most significant bit-plane (MSB), bit-plane coding, 418
Mother wavelet, 122
Motion (camera) matrices, 254–255
Motion blur, 96
Motion coding, 3D-HEVC, 511
Motion compensated prediction. See MCP (motion-compensated prediction)
Motion compensation. See MC (motion compensation)

Motion detection
- motion-adaptive filters using, 345
- motion-adaptive scan-rate conversion and, 365–366
- with successive frame differences, 292

Motion estimation
- 2D. See 2D apparent-motion estimation
- 3D motion/structure. See 3D motion/structure estimation
- 3D-wavelet/sub-band coding benefits, 464
- differential methods, 225–233
- exercises, 268–272
- image formation and, 196–201
- matching methods. See Matching methods, motion estimation
- MATLAB resources, 272
- MC filter in, 347–349
- motion models. See Motion models
- motion segmentation simultaneous with, 313–317
- motion-adaptive filters not needing, 345
- nonlinear optimization methods, 245–249
- overview of, 195–196
- performance measures, 224–225
- transform-domain methods, 249–251

Motion JPEG 2000 (MJ2K), 463

Motion models
- 2D apparent-motion models, 210–214
- apparent motion models, 206–207
- overview of, 202–203
- projected 3D rigid-motion models, 207–210
- projected motion models, 203–206

Motion or object tracking, 274

Motion segmentation
- as change detection. See Change detection, image segmentation
- dominant-motion segmentation, 299–302
- motion estimation simultaneous with, 313–317
- multiple-motion. See Multiple-motion segmentation
- overview of, 298–299
- region-based, 311–313

Motion smoothness, 247–249

Motion snake method, 325–327

Motion tracking
- 2-D mesh tracking, 327–328
- active-contour tracking, 325–327
- graph-based spatio-temporal segmentation, 319
- Kanade–Lucas–Tomasi tracking, 319–321
- mean-shift tracking, 321–322
- overview of, 317–318
- particle-filter tracking, 323–325

Motion trajectory, frequency spectrum of video, 342–343

Motion vector coding, HEVC, 496

Motion vectors. See MVs (motion vectors)

Motion-adaptive de-interlacing, 358

Motion-adaptive filtering, 345

Motion-adaptive noise filtering, 367–369

Motion-adaptive scan-rate conversion, 365–366

Motion-compensated (MC) de-interlacing, 359–361

Motion-compensated (MC) filtering. See MC (motion-compensated) filtering

Motion-compensated multi-frame (MCMF) filter, 377

Motion-compensation error (MCE), 223–224

Motion-field model, 314–315

Motion-picture industry, and Motion JPEG 2000, 463

Motion-skip mode coding tool, MVC, 506

Motion-vector field between two frames, 311–313

Moving Picture Experts Group. See MPEG (Moving Picture Experts Group)

MPC (maximum matching pel count), block matching, 234

MPEG (Moving Picture Experts Group) history of, 467
- inverse pull-down methods in, 363
- video/audio compression with, 86

MPEG HEVC/H.265 standard, 467
MPEG-1 standard
- encoding, 475–476
- input-video format/data structure, 469–471
- inter-frame compression modes, 472–474
- intra-frame compression modes, 471–472
- overview of, 468–469
- quantization and coding, 474
- video compression with, 78

MPEG-2 standard
- for digital broadcast, 467
- encoding, 482–483
  - H.264/AVC vs., 483–484
- input-video format/data structure, 477–478
- interlaced-video compression, 478–480
- other tools and improvements, 480–481
- overview of, 476–477
- profiles and levels, 482
- scalability tools introduced in, 498
- temporal scalability, 498
- video compression with, 78

MPEG-4 AVC and HEVC, 78

MPEG-4 AVC/ITU-T H.264 standard
- See H.264/AVC (MPEG-4 AVC/ITU-T H.264) standard

MPEG-DASH, HTTP-based streaming, 94

MQUAL
- in MPEG-1, 471–472, 475
- in MPEG-2, 481

MRF (Markov random field), adaptive image processing, 153

MS (mean-shift) algorithm, 279–280, 321–322

MSB (most significant bit-plane), bit-plane coding, 418

MSE (mean square difference), 99, 234, 241

MSEA (multi-level SEA), hierarchical block matching, 240

MTU (maximum transfer unit), 484–485

Multicast streaming, 94–95

Multi-dimensional (MD) signals, 1–2, 5–8, 48

Multi-frame noise filtering
- adaptive-weighted-averaging filtering, 372–373
- BM4D filtering, 374
- motion-adaptive noise filtering, 367–369
- motion-compensated noise filtering, 369–372
- overview of, 367
- temporally coherent NLM filtering, 374

Multi-frame restoration, 168, 374–377

Multi-frame SR (super-resolution)
- in frequency domain, 386–389
- limits of, 381
- modeling low-resolution sampling, 381–386
- overview of, 377–378
- recognition/example-based vs. reconstruction-based, 378
- spatial-domain methods, 389–394
- super-resolution vs. image interpolation, 379
- super-resolution vs. image restoration, 379
- what makes SR possible, 379–380
- what super-resolution is, 378

Multi-hypothesis MCP, H.264/AVC, 488

Multi-level SEA (MSEA), hierarchical block matching, 240

Multi-object motion segmentation, 301–302

Multi-picture MCP, H.264/AVC, 487–488

Multiple image displays, 80–81

Multiple motions, 206, 250

Multiple-motion segmentation
- clustering in motion-parameter space, 302–306
- MAP probability segmentation, 309–311
- maximum-likelihood segmentation, 306–309
- overview of, 302

Multi-resolution frame difference analysis, 293

Multi-resolution pyramid representations, 120–121

Multi-resolution representation, wavelet decomposition as, 127

Multi-scale representation, wavelet decomposition as, 127

Multi-view video
- affine reconstruction, 254–255
- compression. See Stereo/multi-view compression
- overview of, 83–85
- projective factorization, 258–259
Multi-view video coding (MVC) standard, 83, 503, 504–507
Multi-view-video-plus-depth (MVD) format, 84, 507
Mumford–Shah functional, 289
Mutual occlusion, 2D motion estimation, 222–223
MVC (multi-view video coding) standard, 83, 503, 504–507
MVC+D (MVC plus depth maps) standard, 507, 509
MVD (multi-view-video-plus-depth) format, 84, 507
MV-HEVC, 508
MVs (motion vectors)
aperture problem and, 222–223
backward extension of, 360
basic block matching, 234
displaced-frame difference method, 219–220
H.264/AVC improvements, 487
HEVC, 496
hierarchical block matching, 240
hierarchical iterative refinement, 226–227
in MPEG-2 encoding, 482–483
multi-frame restoration degraded by PSF, 374–375
in pel-recursive motion estimation, 246
in variable-size block matching, 234, 239–240

NE (norm of the error), motion estimation, 223–224
Near-lossless coding, predictive-coding mode, 428
Negative of image, in linear contrast manipulation, 139–140
Netravali-Robbins algorithm, 246
Network-access layer. See NAL (network-access layer) units
NLM (non-local means) filtering, image denoising, 162–163
Node points, 2D mesh, 327–328
Noise
edge detection sensitivity to, 128
image denoising. See Denoising, image models, 148–149, 152
multi-frame filtering of. See Multi-frame noise filtering
variance in AWA filter, 373
as visual artifact, 96
Noiseless, Lempel–Ziv coding as, 431
Non-interleaved ordering, JPEGs, 436
Non-linear depth representation (NDR), 3D-AVC, 508
Nonlinear filtering
bi-lateral filter, 161–162
image denoising with, 158
in interpolation, 113
median filtering, 158–159
order-statistics filters, 159–160
wavelet shrinkage, 160–161
Nonlinear least-squares problem, bundle adjustment, 259–260
Nonlinear optimization methods, 245–249
Nonlinear wavelet shrinkage, 160
Non-local filtering, 162–164
Non-local image self-similarities, sparse-image modeling, 174
Non-local means (NLM) filtering, image denoising, 162–163
Non-locally centralized sparse representation, image restoration, 174
Non-normative tools, 3D-AVC and MVC+D, 509
Non-parametric model  
estimating 2D motion, 213–214  
Horn–Shunck motion estimation as,  
230–233  
mean-shift (MS) clustering as, 280  
Non-rigid scene, multiple motions with possible  
camera motion, 206  
Non-symmetric half-plane (NSHP) support,  
2–3, 28, 29  
Non-symmetric half-plane symmetry, 5  
Non-uniform quantization, 406  
No-reference metrics (NR), 98–100  
Norm of the error (NE), motion estimation,  
223–224  
Normal flow, OFE, 218–219  
Normalization, 256, 451  
Normalized DLT (normalized 8-point  
algorithm), 243–244  
Normalized rgb, 71–74  
Normalized-cut criterion, graph-based  
segmentation, 286–287  
N-point DCT, 19–20  
NR (no-reference metrics), 98–100  
NSHP (Non-symmetric half-plane) support,  
2–3, 28, 29  
NTSC (National Television Systems Committee)  
analog video format, 64  
ATSC signals using same bandwidth of, 86  
as composite video format, 66  
digitization of analog video from PAL to,  
74–75  
n-view plus n-depth format, 503  
Nyquist criterion, sampling on lattice, 36–41  
Nyquist gain, JPEG2000 filter normalization,  
451  
Nyquist sampling rate, and super-resolution,  
378  

O  
Observation noise, 170, 285  
Occlusions, 85, 221–223, 349  
Oculomotor mechanisms, human stereo vision,  
62–63  

OFE (optical-flow constraint)  
differential methods of, 125–131  
displaced-frame difference and,  
219–220  
overview of, 217–218  
specifying motion field, 218–219  
Open GOPs, HEVC, 491–492  
Optical blurring, 164. See also Image  
restoration  
Optical flow  
dense-motion estimation, 215–216  
estimation problem, 216–219  
Horn–Shunck motion estimation, 231–232  
segmentation. See Motion segmentation  
Optical-flow constraint. See OFE (optical-flow  
constraint)  
Order-statistics filters, 159–160  
Orthogonal filters, 124–127  
Orthogonal sampling, 32–34  
Orthogonal wavelet transform, 160–161  
Orthogonality  
in analysis-synthesis filters, 445–446  
in IIR and FIR LLMSE filters, 150–151  
in image compression, 125  
Orthographic projection model, 199–200  
Otsu method, of threshold determination,  
276–277  
Out-of-focus blurs, point-spread function of,  
165–166  

P  
P2P (peer-to-peer) streaming, 94–95  
P2PTV networks, 95  
Packetized elementary streams (PES), 85–86  
PAL (Phase) format  
analog video format, 64  
as composite video format, 66  
digitization of analog video from NTSC to,  
74–75  
Parallax barriers, auto-stereoscopic multi-view  
displays, 80–81  
Parallel encoding/decoding, HEVC, 493–495  
Parallel projection, 200
Parametric model

- in 2D apparent-motion, 210–213
- of blur identification, 176
- of dominant motion, 300
- of Lukas–Kanade motion estimation, 225–230
- of motion segmentation, 298–299

Paraperspective projection model, 201

Parseval’s Theorem, 18, 19

Partial camera calibration, 253, 260

Partial derivative estimation

- edge detection and image gradient, 127–128
- with finite-differences, 128–131
- with Gaussian filter, 131–132
- hierarchical iterative-refinement, 229
- with Horn–Shunck motion estimation, 231
- Laplacian with finite-difference operators, 133–134

Partial difference equations, and IIR filters, 27–29

Partial differential equations (PDEs), 180–181

Partial derivative estimation

- edge detection and image gradient, 127–128
- with finite-differences, 128–131
- with Gaussian filter, 131–132
- hierarchical iterative-refinement, 229
- with Horn–Shunck motion estimation, 231
- Laplacian with finite-difference operators, 133–134

Periodicity matrix, 37, 48–49

Periodic signals

- finite extent signals isomorphic to, 14
- Fourier series coefficients as, 14
- Fourier transform of discrete signals as
  - rectangularly, 12
  - as isomorphic to finite extent signals, 5
- MD Fourier transform of discrete signals as, 13
  - overview of, 3–4

Photographic film, 64

Photometric effects of 3D motion, 201–202

Photoreceptors, human eye, 54

Picture types

- H.264/AVC not using, 484–485
- in MPEG-1, 469
- in MPEG-2 video, 477

Pixel correspondences, two-view projective reconstruction, 256

Pixel replication, with zero-order-hold filter, 115–116

Pixel sampling density, 68

Pixel-based contrast enhancement

- definition of, 137
- histogram equalization, 140–141

Pearson linear correlation coefficient, 98

Performance

- evaluating segmentation/tracking, 330–331
- limits, in image denoising, 149–150
- motion estimation, 224–225
- quantizer, 406

Perfect-reconstruction (PR) property, 122–123
Index

histogram shaping, 141
image histogram, 138
linear contrast manipulation, 139–140
local contrast manipulation by pixel-based operators, 141–142
overview of, 137–138
Pixel-based motion segmentation, 311–313
Pixel-based operators, contrast enhancement, 141–142
Pixel-difference, change detection, 289–290
Pixel-resolution (spatial) scalability, 499–500, 502
Pixels, bitmap parameters, 68–69
Pixel-wise filtering
bi-lateral filters as, 161–162
median filtering as, 158–159
NLM filtering as, 162–163
nonlinear filters as, 158
order-statistics filters as, 159–160
Planar-parallax
background subtraction with, 296
structure reconstruction, 261–263
PM&S (pattern matching and substitution),
JBIG2 encoder, 423–424
POCS (projections onto convex sets)
formulation, 177–180, 391–394
Point-spread function. See PSF (point-spread function)
Polarization multiplexing, stereoscopic multi-view displays, 80–81
Polarizing filters, stereoscopic 3D glasses, 91
Polyphase implementation
of decimation filters, 112–113
of interpolation, 117–118
Post-processing dilation filtering (PDF), 509
P-pictures
in MPEG-1, 469
in MPEG-1 encoding, 475
in MPEG-1 inter-frame compression, 472–473
in MPEG-2 encoding, 482–483
in MPEG-2 video, 477–479
PR (perfect-reconstruction) property, 122–123
Pre-calibration methods, cameras, 252
Precision, JPEG standard, 434
Precision of segmentation, 274
Prediction, particle-filter motion tracking, 324
Prediction blocks (PBs), 492–493, 495
Prediction units (PUs), HEVC MV coding, 496
Predictive-coding mode, 424, 427–429
Pre-filtering, in A/D conversion, 66–67
Prefix codes, Huffman coding, 410–413
Pre-processing, JPEG2000, 449
Prewitt operator, 130
Primary colors, mixing to create all colors, 56
Probabilistic smoothness constraints, non-parametric model, 213–214
Probability density function (pdf), 138, 247–249
Probability distribution, of symbols, 402–403
Processing order, H.264/AVC, 485
Profile Connection Space (PCS), 57
Profiles
H.264/AVC, 484
HEVC, 497
MPEG-2 video, 476–477, 482
Program stream (PS), MPEG, 86
Progressive (non-interlaced) video, 469, 477–478
Progressive conversion, interlaced to, 46
Progressive mode JPEG, 441–442
Progressive scanning, 64–65
Projected 3D rigid-motion models, 207–210
Projected motion models, 203–206
Projection slice theorem, 11–12
Projections onto convex sets (POCS)
formulation, 177–180, 391–394
Projective camera, in motion estimation, 196–199
Projective factorization, 258–259
Projective reconstruction
bundle adjustment, 259–260
multi-view projective factorization, 258–259
overview of, 255
two-view epipolar geometry, 255–257
Index

Rational factor, sampling rate change by, 117–118
117–118
Raw (uncompressed) data rates, digital video,
77–78
RCT (reversible color transform), JPEG2000, 449
Real part, frequency response, 24
RealD 3D cinema technology, 91
Real-Time Messaging Protocol (RTMP), 93
Real-time performance, segmentation method
and, 274
Real-Time Streaming Protocol (RTSP), 93
Real-time Transport Protocol (RTP), 93
Real-valued functions, frequency response, 24
Real-valued signals, 14–18
Reciprocal lattice, 35
Recognition/example-based methods, super-
resolution, 378
Reconstruction
filtering, 121
from samples on lattice, 41–42
super-resolution methods, 378
Rectangular periodic signal, 2D, 4–5, 12
Rectangular periodicity, 2D, 16
Rectangular sampling, 2D, 37, 49
Recursive filters, 28–29
Recursively computable prediction model, 424
Red, green, blue. See RGB (red, green, blue)
model
Reduced reference metrics (RR), 98–100
Reduced-resolution depth coding, 3D-AVC,
508–509
Redundancy reduction, 405
Reference element, two-dimensional RLC,
421–423
References
digital images and video, 100–103
image compression, 454–456, 459
image filtering, 181–186
MD signals and systems, 47–48
motion estimation, 263–268
video compression, 512–514, 515
video segmentation and tracking, 331–338
Reflection, Lambertian reflectance model,
201–202
Refresh rate, 61, 65, 76
Region-based motion segmentation, 311–313
Region-of-interest (ROI) coding, 454, 502
Regular mode, JPEG-LS, 426–427
Regularization
operator choices, CLS, 173
in restoring images from LSI blur, 170–173
by sparse-image modeling, 173–174
Regularized deconvolution methods, 170–173
Relative address coding, two-dimensional RLC,
421–423
Relative affine structure reconstruction, 263
Re-sampling
decimation, 111–113
gradient estimation. See Gradient estimation/
edge/features
interpolation, 113–120
multi-resolution pyramid representations, 120
multi-resolution wavelet representations,
121–127
overview of, 110–111
in particle-filter motion tracking, 324
Residual planar-parallax motion model,
projected 3D rigid-motion, 209–210
Resolution
independence of JPEG standard, 434
multi-frame super. See Multi-frame SR (super-
resolution)
reconstruction from samples on lattice, 41–42
sampling structure conversion for, 42–47
super-resolution. See Multi-frame SR (super-
resolution)
of volumetric display, 82
Restoration, image
blind restoration/blur identification, 175–176
blur models, 165–168
boundary problem in, 175
degradation from linear space-invariant blurs,
168–174
degradation from space-varying blurs,
177–180
image in-painting for, 180–181
overview of, 164–165
super-resolution vs., 379
Restoration, multi-frame video
- cross-correlated multi-frame filter, 377
- MC multi-frame filter, 377
- multi-frame modeling, 375
- multi-frame Wiener restoration, 375–377
  overview of, 374–375
Retina, of human eye, 54
Retinex, 146, 147
Retouching, image-in-painting taken from,
  180
Reversible color transform (RCT), JPEG2000, 449
RFID (radio frequency identification) motion tracking, 317
RGB (red, green, blue) model
  in color image processing, 71–73, 105
  color management, 57
  in component analog video, 65–66
digital color cameras and, 67–68
hue-saturation-intensity, 73–74
human eye processing of, 54–56
JPEGs, 436
  normalized rgb, 73
  overview of, 56
  three-sensor cameras capturing, 69
Rigid scene, projected motion with static camera in, 204–205
Rigid-motion models, projected 3D, 207–210
Ringing artifacts, 97, 442–443
RLC (run-length coding), 419–423
Roberts cross operator, gradient estimation, 130–131
Rods, sensitivity of retinal, 54
ROI (region-of-interest) coding, 454, 502
Rotation matrix, in perspective projection, 199
RR (reduced reference metrics), 98–100
RTMP (Real-Time Messaging Protocol), 93
RTP (Real-time Transport Protocol), 93
RTSP (Real-Time Streaming Protocol), 93
Run mode, JPEG-LS, 426–427
Run-length coding mode, JPEG-LS, 429–430
Run-length coding (RLC), 419–423
S
S (short) wavelength, retinal cone sensitivity to, 54
Saccadic eye movements, 62
SAD (sum of absolute differences), 236–240, 512
Sample preservation property, interpolation filters, 115
Sampling
  3D sampling on lattices, 33–34, 50
  in analog-to-digital conversion, 66–67
derivatives of Gaussian filtering, 131–132
  with Fourier transform to obtain DFT, 14
  low-resolution, 381–386
  MD theory of. See MD (multi-dimensional)
sampling theory
  rate change by rational factor, 117–118
Sampling matrix, 31
Sampling rate
  in aliasing, 40
  in analog-to-digital conversion, 67
causes of blur, 96
  in CD-quality digital audio, 71
  change by rational factor, 117–120
  frame/field rate-up conversion increasing temporal, 350
  in image decimation, 112
  in multi-rate digital signal processing, 111
  in super-resolution, 378
  what super-revolution is, 378
Sampling structure conversion, 350
Satellite television standards, 86–88
Saturation, HSI and, 73–74
Scalability
  of 3D-wavelet/sub-band coding, 464–466
  pixel-resolution (spatial), 499–500, 502
  SVC. See SVC (scalable-video coding)
    compression
      temporal, 498
      in wavelet analysis, 122
Scalar quantization, 406–409, 432
Index

Scale-invariant feature transform (SIFT) system, 137
Scanning, progressive vs. interlaced, 64–65
Scan-rate doubling conversion, 363–365
SDTV (standard definition TV), 75, 78
SEA (successive elimination algorithm), 236, 240
SECAM (Systeme Electronique Color Avec Memoire), 64, 66
Second derivatives, edge of 1D continuous signal, 127–128
Sectional processing, space-varying restorations, 177
Segmentation. See Video segmentation and tracking
SEI messages, in H.264/AVC and /H.265/HEVC, 504
Self-occlusion, 2D motion estimation, 222–223
Semantically meaningful object segmentation, 328
Separable filters, 22–23
Separable signals, MD, 6
Sequences, MPEG-1, 469
Sequential Karhunen–Loeve (SKL) algorithm, 325
Server-client streaming, 92–93
Set-theoretic methods, 391–394
SFM (structure from motion) methods, 251–252. See also 3D motion/structure estimation
SFS (structure from stereo), 251, 263
Shaping, histogram, 141
Sharpening image. See Spatial filtering
Shift of Origin, properties of DFT, 17–18
Shift-varying spatial blurring, 168
Shi–Tomasi corner detection method, 136
Short (S) wavelength, retinal cone sensitivity to, 54
Shot-boundary detection, 289–291
SHVC, HEVC extended to, 498
SI (switching I) slice, H.264/AVC, 485
SIF (standard input format), 469
SIFT (scale-invariant feature transform), 137
Signal formats, analog video, 65–66
Signal-dependent noise, 148
Signal-independent noise, 148
Signals. See MD (multi-dimensional) signals
Signal-to-noise ratio (SNR), 149, 482
Similarity transformation, affine model, 212
Simple profile, MPEG-2, 482
Simulated annealing, MAP segmentation, 283
Simulation model encoder (SM3), MPEG-1, 475–476
Simultaneous contrast effect, 58
Simultaneous motion estimation/segmentation, 313–317
Sinc function, ideal interpolation filter, 115
Singular-value decomposition (SVD), 254–255
Skip frame, video transmission on unreliable channels, 97
SKL (sequential Karhunen–Loeve) algorithm, 325
Slices
H.264/AVC, 484–486
MPEG-1, 469
SM3 (simulation model encoder), MPEG-1, 475–476
Smearing. See Blur
Smooth pursuit eye movements, 62
Smooth Streaming, Microsoft, 94
Smoothing, image
bi-lateral filtering for, 109–110
box filtering for, 107–108
Gaussian filtering for, 108–109
linear, shift-invariant low-pass filtering for, 106–109
overview of, 106
SMPTES (Society of Motion Picture and Television Engineers), 66–67, 74
SMV (super multi-view) displays, 81–82
Snake method, video matting, 329
Snakes (active-contour models), 287–289
SNR (signal-to-noise ratio), 149, 482
Sobel operator, gradient estimation, 130
Source encoder, 405
Source encoding, 405
Source-coding theorem, 404–405
SP (switching P) slice, H.264/AVC, 485
Space-frequency spectral methods, motion estimation, 251
Space-invariant spatial blurs, 167
Space-varying blurs, 177–180
Space-varying class-means, 284–285
space-varying image model, 155–156
Sparse feature matching, motion estimation, 234
Sparse modeling, single-frame SR, 119
Sparse priors, in-painting problem, 181
Sparse representations, for denoising, 160–161
Sparse-correspondence estimation, 2D, 214
Spatial (pixel-resolution) scalability, 499–500, 502
Spatial artifacts, 96
Spatial consistency, in background subtraction, 295
Spatial filtering
  adaptive filtering, 145–146
  definition of, 137
  digital dodging-and-burning, 142–143
  homomorphic filtering, 146–147
  image enhancement, 142
  retinex, 146
  unsharp masking, 144–145
Spatial masking, 59
Spatial partial, 229, 231
Spatial profile, MPEG-2, 482
Spatial redundancy, and image compression, 401
Spatial resolution formats, digital-video, 67–69
Spatial resolution (picture size)
in auto-stereoscopic multi-view displays, 81
frame-compatible formats losing, 503
of video format, 349
Spatial segmentation, background subtraction with, 295
Spatial weighting, hierarchical iterative-refinement, 229
Spatial-domain aliasing, 14, 28
Spatial-domain methods, multi-frame, 389–394
Spatial-frequency patterns
  blur due to, 96
  Fourier transform of continuous signals and, 8
  MD signals and, 6
  in spectrum of signals stamped on lattice, 35
Spatial-frequency response, human vision, 60–62
Spatio-temporal filtering
  frequency spectrum of video, 342–345
  motion-adaptive filtering, 345
  motion-compensated filtering, 345–349
  overview of, 342
Spatio-temporal intensity pattern, video as, 53
Spatio-temporal median filtering, 365–366
Spatio-temporal resolution
  interlacing and, 47
  reconstruction from samples on lattice, 41–42
  sampling structure conversion for, 42–47
Spatio-temporal segmentation, 274
Spearman rank-order correlation coefficient, 98
Special multidimensional signals, 5
Spectral redundancy, 401
Spectrum of signals, sampled on lattice, 34–36
SPIHT (set partitioning in hierarchical trees), 448
SPM (soft pattern matching) method, 423–424
Sports video, shot-boundary detection challenges, 290
SR (super-resolution). See Multi-frame SR (super-resolution)
sRGB (CIEXYZ) color space, 56, 57
SSD (sum of squared differences), 135–136, 512
SSIM (structural similarity index), 99
Stability, testing for 2D recursive filtering, 28
Stable filters, 21–22
Standard definition TV (SDTV), 75, 78
Standard input format (SIF), 469
Standards
  adaptive streaming, 94
  analog-video, 65–66
Index

digital cinema, 89–90

digital-video, 73–74

image/video compression, 78

sampling parameter, 66–67

sampling structure conversion, 42–47

subjective quality assessments, 97–98

Static camera, projected motion with, 203–204

Static scene, projected motion in, 203–204

Static-volume 3D displays, 82

Statistical redundancy, data compression by, 405

Stein's unbiased risk estimate (SURE), 161

Step-size, uniform quantizer, 407

Stereo, dense structure from, 263

Stereo vision, 62–63, 203–204

Stereo/depth perception, 62–63

Stereo-disparity estimation, 216

Stereo/multi-view compression

frame-compatible formats, 503–504

MVD extensions, 507–512

overview of, 502–503

video coding extensions of H.264/AVC, 504–507

Stereoscopic (with glasses), 80–83, 91, 100

Stereoscopy, creating illusion of depth, 62–63

Stereovideo information (SVI) messages, H.264/AVC, 504

Still images, 53, 60

Storage, MPEG-2 built for DVD, 78

Stream switching, 92

Structural similarity index (SSIM), 99

Structure from motion (SFM) methods, 251–252. See also 3D motion/structure estimation

Structure from stereo (SFS), 251, 263

Sturm–Triggs method, 259

Sub-band coding

3D-wavelet and, 464–466

DWT and, 443–447

wavelet image coding vs., 447–448

wavelet-transform coding and, 443

Sub-pixel displacement, for super-resolution, 379–380

Sub-pixel motion estimation, SR without, 394

Sub-pixel search, block-matching, 238

Sub-sampling. See Down-sampling (sub-sampling)

Subtractive color model, 56–57

Successive elimination algorithm (SEA), 236, 240

Successive iterations, regularization by CLS, 172–173

Sum of absolute differences (SAD), 236–240, 512

Sum of squared differences (SSD), 135–136, 512

Super multi-view (SMV) displays, 81–82

Super multi-view video, 83–85

Super voxels, graph-based video segmentation, 319

Super-resolution. See Multi-frame SR (super-resolution)

Suppression theory of stereo vision, 63

SURE (Stein's unbiased risk estimate), 161

SureShrink, wavelets, 161

SVC (scalable-video coding) compression

benefits of, 497–498

hybrid scalability, 502

quality (SNR) scalability, 500–502

spatial scalability, 499–500

temporal scalability, 498–499

SVC (scalable-video encoding) compression, 92

SVD (singular-value decomposition), 254–255

SVI (stereo-video information) messages, H.264/AVC, 504

S-video (Y/C video), formats, 65–66

Sweet spots, auto-stereoscopic multi-view displays, 81

Swept-volume 3D displays, 82

Switching I (SI) slice, H.264/AVC, 485

Symbols

coding in image compression, 409–410

information content of, 402–403

Symlet filters, 124–125

Symmetric block-matching, MC up-sampling, 348

Symmetric boundary extension, JPEG2000, 451

Symmetric filters, 26–27, 125–127
Symmetric signal, 5, 20
Symmetry
  analysis-synthesis filtering property, 123
  in analysis-synthesis filters, 445
  as DCT property, 18–19
  as DFT property, 16
  FIR filters in MD systems and, 25–27
  MD Fourier transform of discrete signals as, 13
  properly handling image borders with, 125
Synthesis filters, wavelet analysis, 122–124
Systeme Electronique Color Avec Memoire (SECAM), 64, 66
Systems, MPEG-2, 476

T
TBs (transform blocks), HEVC, 493, 495
TCP (transmission control protocol), server-client streaming, 92–93
Television. See TV (television)
Template-tracking methods, 318, 321
Temporal (inter-frame) redundancies, 461, 462
Temporal artifacts, 96
Temporal consistency, 295, 301
Temporal masking, 59
Temporal partial, 229, 231
Temporal prediction, MPEG-1, 472
Temporal scalability, 498–499, 502
Temporal segmentation, 289
Temporal-frequency, 61–62, 342–343
Temporally coherent NLM filtering, 374
Temporal-prediction structures, H.264/AVC, 485
Terrestrial broadcast standards, 86–87
Terrestrial mobile TV broadcasts, DVB-H for, 89
Test Model 5 (TM5) encoder, MPEG-2, 482–483
Test Zone Search (TZZearch) algorithm, 239–240
Testing, and subjective quality assessments, 97–98
Texture coding tools, 3D-AVC, 509
Texture mapping, 2D-mesh tracking, 327–328
Three-step search (TSS), matching method, 236–237
Thresholding
  determining edge pixels by, 134
  estimating in wavelet shrinking, 160–161
  finding optimum threshold, 276–277
  for image segmentation, 275–276
  in multi-resolution frame difference analysis, 293
Tier-1 coding, JPEG 2000 operation, 448
Tier-2 coding, JPEG 2000 operation, 448
Tiers, HEVC, 497
Tiles, HEVC, 493–494
Tiles, JPEG2000, 448, 449
Time-first decoding, MVC, 505–506
Time-recursive (TR) filters, MC de-interlacing, 360–361
Time-sequential sampling, 33–34
TM5 (Test Model 5) encoder, MPEG-2, 482–483
Tone mapping. See Pixel-based contrast enhancement; Spatial filtering
TR (time-recursive) filters, MC de-interlacing, 360–361
Transform blocks (TBs), HEVC, 493, 495
Transform coding
  DCT. See DCT (discrete cosine transform)
  DCT vs. H.264/AVC, 488–489
  H.264/AVC, 488–489
  HEVC, 496
  MD Fourier transform. See MD (multi-dimensional) transforms
Transform-domain methods, motion estimation, 249–251
Transmission control protocol (TCP), server-client streaming, 92–93
Transport stream, ATSC, 86–87
Transport stream (TS), MPEG, 86
Tree diagram, Huffman coding, 412
Triangulation, projective reconstruction, 257
Trimap (tri-level image segmentation), 328
Tri-stimulus theory, digital color, 56
Tri-stimulus values, color vision and, 54
TS (transport stream), MPEG, 86
TSS (three-step search), matching method, 236–237
Index

TV (television)
  analog, 64
  broadcast standards, 74–76
  digital TV (DTV), 85–89
  frame rate for, 61
  frame rate standards, 361
  frame-compatible stereo-video formats for
    3DTV, 83
  intra-frame compression in, 462
  scan-rate doubling used in, 363
  Two-field median filter, motion-adaptive
de-interlacing, 358
  Two-fold symmetry
    definition of, 5
    in Fourier domain, 14
    impulse response of low-pass filter, 27
    MD Fourier transform of discrete
      signals, 14
    signals, 5
  Two-level binary-tree decomposition, 126
  Two-step iteration algorithm, 316–317
  Two-view epipolar geometry, projective
    reconstruction, 255–257
  Types I-IV DCT, 18–19
  Types V-VIII DCT, 18
  TZZearch (Test Zone Search) algorithm,
    239–240

U
  UDP (User Datagram Protocol), 92–93, 95
  UHDTV (ultra-high definition television),
    76, 78
  Ultra-high definition television (UHDTV),
    76, 78
  UMHexagonS, matching method, 237–238
  Uncompressed (raw) data rates, digital video,
    77–78
  Uncovered background, in motion estimation,
    222–223
  Uniform convergence, 9, 12
  Uniform quantization
    defined, 406
    in JPEG2000, 448, 451–452
    with Lloyd-Max quantizers, 407
    in MPEG-1, 471
    overview of, 407–409
  Uniform reconstruction quantization (URQ),
    HEVC, 496
  Uniform velocity global motion, 344
  Unit step, MD signals, 7–8
  Unitless coordinates, 32
  Unit-step function, modeling image edges, 127
  Unsharp masking (USM), image enhancement,
    144–145
  Up-conversion, sampling structure conversion,
    43–45
  URQ (uniform reconstruction quantization),
    HEVC, 496
  User Datagram Protocol (UDP), 92–93, 95
  USM (unsharp masking), image enhancement,
    144–145

V
  Variable-block-size MCP, H.264/AVC, 487
  Variable-length source coding. See VLC
    (variable-length source coding)
  Variable-size block matching (VSBM), 234,
    238–240
  Variance of noise, SNR, 149
  VBV (video buffer verifier), MPEG-1, 476
  VCEG (Video Coding Experts Group), ITU-T,
    467
  Vector quantization, 406
  Vector-field image segmentation, 285
  Vector-matrix model, 200, 375
  Vectors
    2D sampling lattices, 32
    3D sampling lattices, 33–34
    MD sampling lattice, 30–31
  Vergence distances, human stereo vision, 62–63
  Vertical mode, two-dimensional RLC, 421–423
  Vertical resolution, video signals, 65
  Vertical spatial-frequency pattern, MD
    signals, 6
  Vertical sum buffer (VSB), box filtering,
    107–108
  VESA (Video Electronics Standards
    Association), 76
VGA (Video Graphics Array) display standard, 76
ViBe algorithm, 295–297
Video
MD signals/systems in, 1
as spatio-temporal intensity pattern, 53, 61–62
temporal-frequency response in, 61
Video buffer verifier (VBV), MPEG-1, 476
Video Coding Experts Group (VCEG), ITU-T, 467
Video compression
3D-transform coding, 463–466
digital TV, 85–86
fast search, 236–238
intra-frame, 462–463
motion-compensated transform coding, 466
overview of, 461–462
scalable video compression, 497–502
stereo/multi-view, 502–512
Video compression standards
digital-video, 74–78
HEVC, 491–497
international, 467
ISO and ITU, 467–468
Motion JPEG 2000, 463
MPEG-1, 468–476
MPEG-2, 476–483
MPEG-4/ITU-T H.264, 483–491
Video Electronics Standards Association (VESA), 76
Video filtering
multi-frame noise filtering, 367–374
multi-frame restoration, 374–377
multi-frame SR. See Multi-frame SR (super-resolution)
overview of, 341
spatio-temporal filtering theory, 342–349
video-format conversion. See Video-format conversion
Video Graphics Array (VGA) display standard, 76
Video matting methods, 329
Video Quality Experts Group (VQEG), 100
Video segmentation and tracking
change detection. See Change detection,
image segmentation exercises, 239
factors affecting choice of method for, 274
image and video matting, 328–329
image segmentation. See Image segmentation
motion segmentation. See Motion segmentation
motion tracking. See Motion tracking
overview of, 273–275
performance evaluation, 330–331
Video streaming over Internet, 92–95
Video view coding, 3D-HEVC, 510
Video-format conversion
de-interlacing, 355–361
down-conversion, 351–355
frame-rate conversion, 361–367
overview of, 349–350
problems of, 350–351
Video-format standards, digital video, 74–77
Video/image quality
objective assessment of, 98–100
overview of, 96
subjective assessment of, 97–98
visual artifacts, 96–97
Video-interface standards, digital video, 77
Video-on-demand request (pull application), 92
View synthesis, 3D-HEVC, 510
View synthesis distortion (VSD), 509
View synthesis prediction coding tool, MVC, 506
View synthesis prediction (VSP), 509, 511
View-first decoding, MVC, 505–506
Viewing distance, human eye response, 60
Viewing-position-dependent effects, volumetric displays, 82
View-plus-depth format, in multi-view video, 84–85
Vision persistence, 61
Visual artifacts, 96–97
Visual masking, human vision, 58–59
Visual motion tracking, 317–318
Index

Visual-quality degradation (loss), image compression, 401
VisuShrink, 161
VLC (variable-length source coding)
  defined, 402
  as entropy coding, 410
  JPEG baseline mode, 435–436
  in MPEG-1, 468, 471–472, 474–475
Volumetric displays, 80, 82
Voronoi cell of a 2D lattice, 36
Voxels, 82
VQEG (Video Quality Experts Group), 100
VQM, objective quality assessment, 99–100
VSB (vertical sum buffer), box filtering, 107–108
VSBM (variable-size block matching), 234, 238–240
VSD (view synthesis distortion), 509
VSP (view synthesis prediction), 509, 511

W
Warping, hierarchical iterative-refinement, 229
Wavefront parallel processing, HEVC, 494–495
Wavelengths, color sensitivity and, 54–56
Wavelet filters, JPEG2000, 450
Wavelet representations
  with bi-orthogonal filters, 125–127
  in image re-sampling, 121–124
  with orthogonal filters, 124–125
Wavelet shrinkage, image denoising, 160–161
Wavelet transform coding
  choice of filters, 443–447
  overview of, 443
  sub-band vs. wavelet image coding, 447–448
  wavelet compression, 448
Weak perspective projection model, 200
Weave filtering (inter-field temporal)
  deinterlacing, 357–358
Weber’s law, 58
Wedge support, MD signals, 2–3
Weighted MCP, H.264/AVC, 488
Weighted median filtering, image denoising, 159
Weights for patches, NL-means filtering, 162–163
White noise
  adaptive LMMSE filter and, 155–156
  defined, 148
  IIR Weiner filter and, 153
  image denoising with wavelet shrinkage, 160
Wiener filter
  collaborative, 164
  deconvolution filter, 170–171, 173
  IIR (infinite-impulse response), 151–153
  image restoration with, 169
  multi-frame restoration, 375–377
  Wiener-based motion estimation, 147

X
x264 library, H.264/AVC, 491
x265 library, HEVC, 497
XGA (Extended Graphics Array), 76
XpanD 3D cinema technology, 91

Y
Y/C video (S-video), formats, 65–66
Y-Cr-Cb color space
  in color image processing, 71–74, 105
  JPEG, 436
  MPEG-1, 469

Z
Zero Fourier-phase, 14
Zero-crossings
  blur identification from, 175–176
  linear motion blur, 167
  multi-frame restoration and, 275
  out-of-focus blur, 166
  super-resolution in frequency domain and, 288
Zero-mean noise, 155–156
Zero-order coder, 426
Zero-order hold filtering, 115–116, 360–361
Zero-phase filters, 25–27, 445
Zigzag scanning
  in MPEG-2, 480
  of quantized AC coefficients, 439, 472
  of quantized DCT coefficients, 435, 438, 441, 474