SECOND EDITION

PRENTICE HALL

DIGITAL VIDEO PROCESSING

PRENTICE HALL SIGNAL PROCESSING SERIES

A. MURAT TEKALP

FREE SAMPLE CHAPTER

SHARE WITH OTHERS

Digital Video Processing

Second Edition

This page intentionally left blank

Digital Video Processing

Second Edition

A. Murat Tekalp

PRENTICE HALL

New York • Boston • Indianapolis • San Francisco
Toronto • Montreal • London • Munich • Paris • Madrid
Capetown • Sydney • Tokyo • Singapore • Mexico City

For information about buying this title in bulk quantities, or for special sales opportunities (which may include electronic versions; custom cover designs; and content particular to your business, training goals, marketing focus, or branding interests), please contact our corporate sales department at corpsales@pearsoned.com or (800) 382-3419.

For government sales inquiries, please contact governmentsales@pearsoned.com.

For questions about sales outside the United States, please contact international@pearsoned.com.

Visit us on the Web: informit.com/ph

Library of Congress Cataloging-in-Publication Data

Tekalp, A. Murat.
Digital video processing / A. Murat Tekalp.—Second edition.
pages cm
Includes bibliographical references and index.
ISBN 978-0-13-399100-0 (hardcover : alk. paper)—ISBN 0-13-399100-8 (hardcover : alk. paper)
1. Digital video—Textbooks. I. Title.
TK6680.5.T45 2015
621.388’33—dc23
2015007504

ISBN-13: 978-0-13-399100-0
ISBN-10: 0-13-399100-8

Text printed in the United States on recycled paper at Courier in Westford, Massachusetts.
First printing, June 2015

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

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 Hence, the first edition contained some now-outdated methods/algorithms and technologies compared with the state of the art today, and obviously missed important developments in the last 20 years.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 nonlinear 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

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

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

About the Author

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 *Proceeedings 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

CHAPTER 2

Digital Images and Video

Advances in ultra-high-definition and 3D-video technologies as well as high-speed Internet and mobile computing have led to the introduction of new video services.

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.

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 λ dimension. Hence, color sensation can be achieved by sampling λ into three levels to emulate color sensation of each type of cones as:

$$f_k(x_1, x_2) = \int f(x_1, x_2, \lambda) m_k(\lambda) d\lambda \quad k = r, g, b \tag{2.1}$$

where $m_k(\lambda)$ is the wavelength sensitivity function (also known as the color-matching function) of the kth 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 color-difference (chrominance) values [Fer 01]. The luminance Y is related to the perceived brightness of the light and is given by

$$Y(x_1, x_2) = \int f(x_1, x_2, \lambda)\, l(\lambda)\, d\lambda \tag{2.2}$$

where $l(\lambda)$ 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.

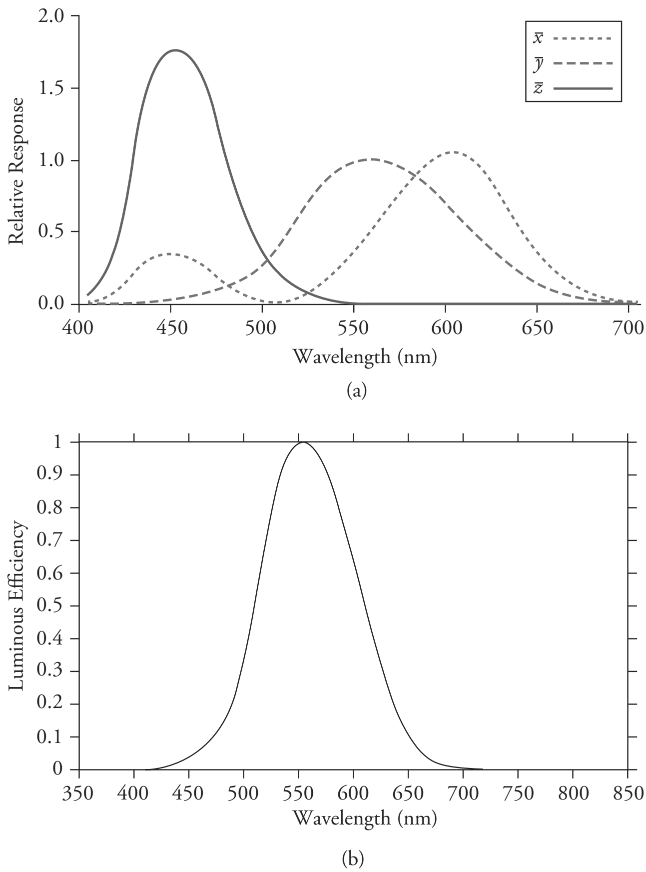

Figure 2.1 Spectral sensitivity: (a) CIE 1931 color-matching functions for a standard observer with a 2-degree field of view, where the curves $\bar{x}$, $\bar{y}$, and $\bar{z}$ may represent $m_r(\lambda)$, $m_g(\lambda)$, and $m_b(\lambda)$, respectively, and (b) the CIE luminous efficiency function $l(\lambda)$ as a function of wavelength λ.

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 CIERGB, CIEXYZ, or sRGB. CIERGB and CIEXYZ 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 limitiation 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.

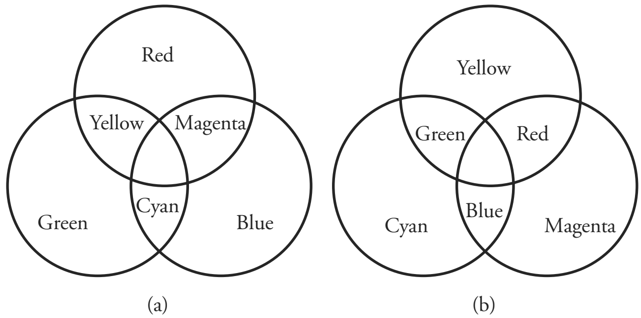

Figure 2.2 Color spaces: (a) additive color space and (b) subtractive color space.

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

$$\frac{\Delta I}{I} = c \text{ (constant)}, \text{ for } I > 0 \tag{2.5}$$

where ΔI 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 ΔI increases relative to the case $I_0 = I$. Furthermore, the *simultanenous 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

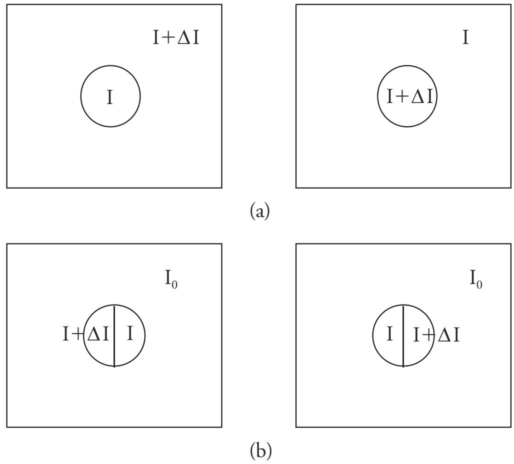

Figure 2.3 Illustration of (a) the just noticeable difference and (b) brightness adaptation.

brightness or failure to detect the target or some details, e.g., texture. Visual masking can be studied under two cases: *spatial masking* and *temporal masking*.

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 θ 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,

$$\theta \approx \frac{w}{d}(\text{radians}) = \frac{180w}{\pi d}(\text{degrees}) \tag{2.3}$$

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_θ is given by

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

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_θ 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.

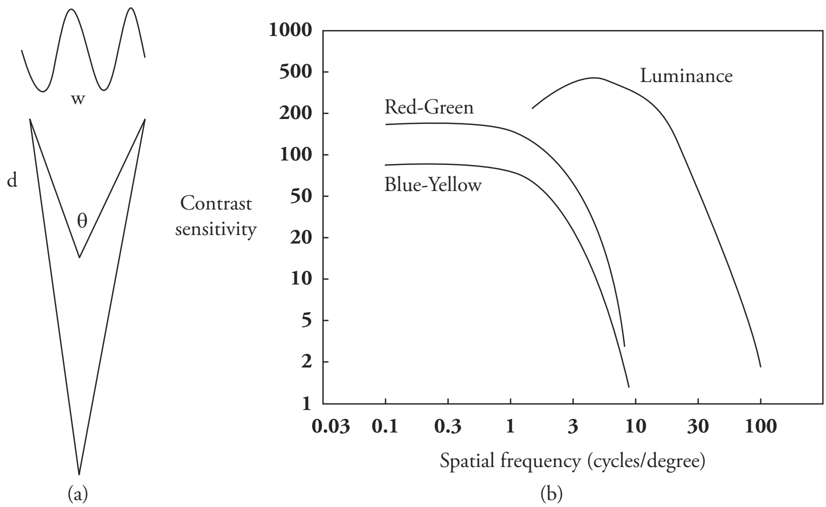

Figure 2.4 Spatial frequency and spatial response: (a) viewing angle and (b) spatial-frequency response of the human eye [Mul 85].

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 *interpupilar distance*) of an average human is 6.5 cm. The difference between the left and right retinal images is called *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.

Accomodation, 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.

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 $s_c(x_1, x_2, t)$ 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 $s_c(x_1, x_2, t)$ 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 Δt 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 Δt=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

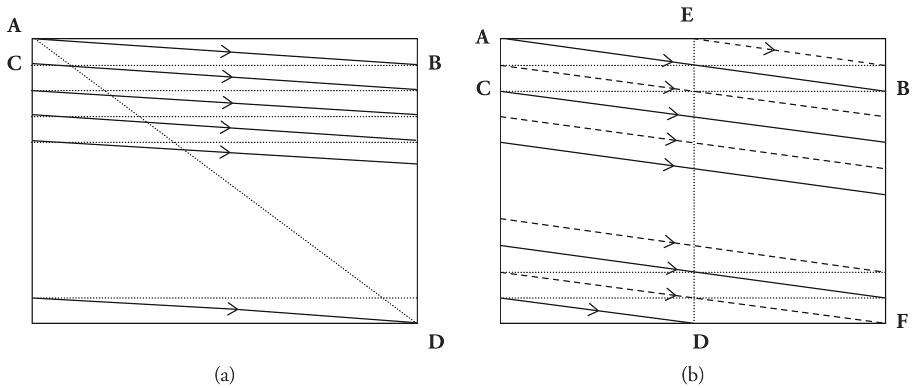

Figure 2.5 Scanning raster: (a) progressive scan; (b) interlaced scan.

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

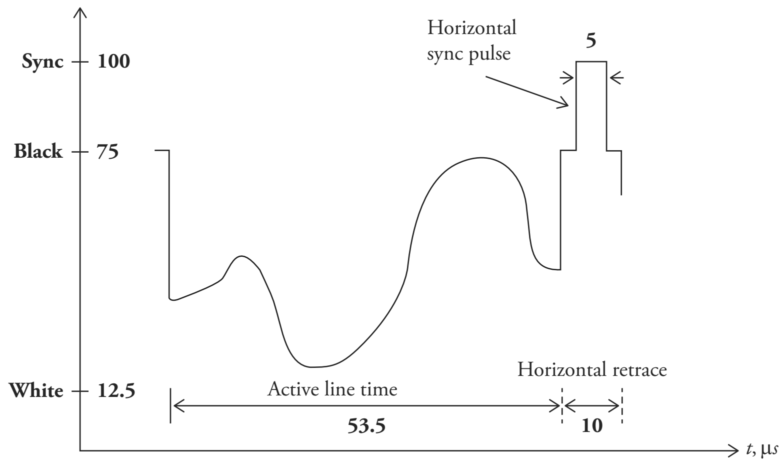

Figure 2.6 Analog-video signal for one full line.

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 videocasette 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

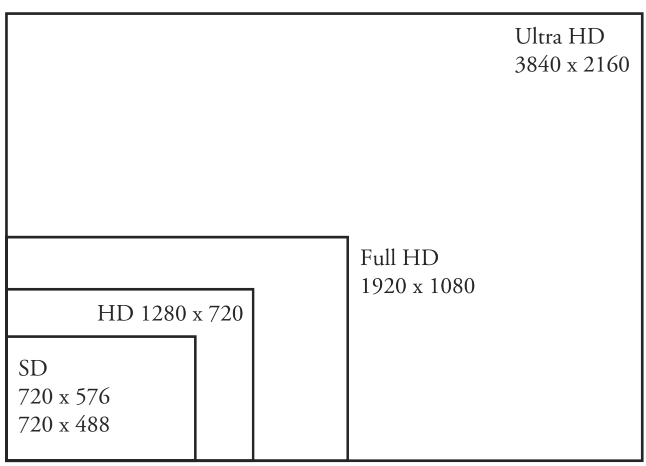

Figure 2.7 Digital-video spatial-resolution formats.

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

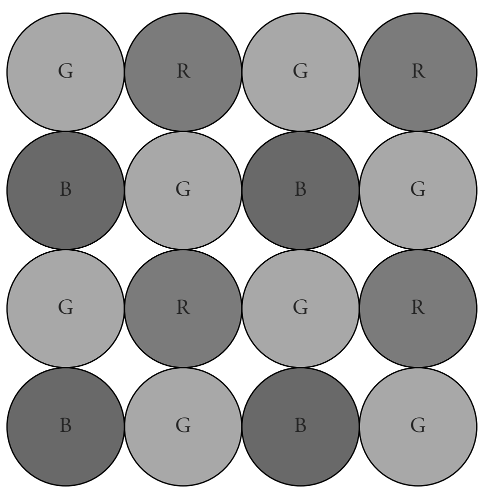

Figure 2.8 Bayer color-filter array pattern.

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 cd/m^2 (starlight) to 6 log cd/m^2 (sun light); 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/m^2, compared to standard displays with contrast ratio 2 log units and maximum intensity 300 cd/m^2, 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}

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.

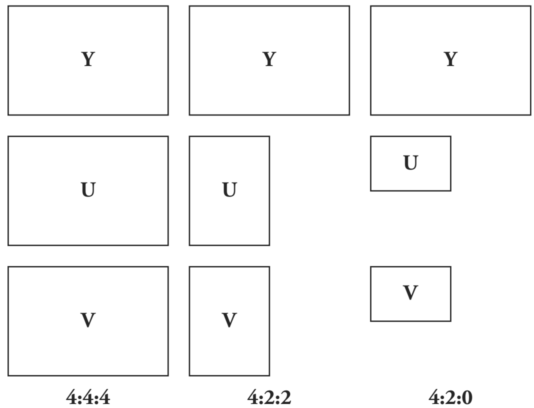

Figure 2.9 Chrominance subsampling formats: (a) no subsampling; (b) 4:2:2; (c) 4:2:0 format.

ITU-R BT.709 defines the conversion between RGB and YCrCb representations as:

$$\begin{aligned} Y &= 0.299\ R + 0.587\ G + 0.114\ B \\ Cr &= 0.499\ R - 0.418\ G - 0.0813\ B + 128 \\ Cb &= -0.169\ R - 0.331\ G + 0.499\ B + 128 \end{aligned} \tag{2.6}$$

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

$$\begin{aligned} R &= Y + 1.402\ (Cr - 128) \\ G &= Y - 0.714\ (Cr - 128) - 0.344\ (Cb - 128) \\ B &= Y + 1.772\ (Cb - 128) \end{aligned} \tag{2.7}$$

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 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

$$\begin{aligned} r &= R/(R+G+B) \\ g &= G/(R+G+B) \\ b &= B/(R+G+B) \end{aligned} \tag{2.8}$$

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

$$r + g + b = 1 \tag{2.9}$$

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].

$$
\begin{aligned}
H &= \begin{cases} \theta & \text{if } B \leq G \\ 360-\theta & \text{if } B > G \end{cases} \quad \text{where } \theta = \arccos\left\{ \frac{\frac{1}{2}[(R-G)+(R-B)]}{\sqrt{(R-G)^2+(R-B)(G-B)}} \right\} \\
S &= 1 - \frac{3\min\{R,G,B\}}{R+G+B} \qquad (2.10) \\
I &= \frac{R+G+B}{3}
\end{aligned}
$$

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) $f_{h,525} = 525 = 29.97 = 15{,}734$ 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

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

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

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

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

Table 2.1 ITU-R TV Broadcast Standards

Standard	Pixels	Lines	Interlace/Progressive, Picture Rate	Aspect Ratio
BT.601-7 480i	720	486	2:1 Interlace, 30 Hz (60 fields/s)	4:3, 16:9
BT.601-7 576i	720	576	2:1 Interlace, 25 Hz (50 fields/s)	4:3, 16:9
BT.709-5 720p	1280	720	Progressive, 50 Hz, 60 Hz	16:9
BT.709-5 1080i	1920	1080	2:1 Interlace, 25 Hz, 30 Hz	16:9
BT.709-5 1080p	1920	1080	Progressive	16:9
BT.2020 2160p	3840	2160	Progressive	16:9
BT.2020 4320p	7680	4320	Progressive	16:9

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 up-scaling 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 2.2 Display Standards

Standard	Pixels	Lines	Aspect Ratio
VGA	640	480	4:3
WSVGA	1024	576	16:9
XGA	1024	768	4:3
WXGA	1366	768	16:9
SXGA	1280	1024	5:4
UXGA	1600	1200	4:3
FHD	1920	1080	16:9
WUXGA	1920	1200	16:10
HXGA	4096	3072	4:3
WQUXGA	3840	2400	16:10
WHUXGA	7680	4800	16:10

"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 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 2.3 International Standards for Image/Video Compression

Standard	Application
ITU-T (formerly CCITT) G3/G4	FAX, Binary images
ISO JBIG	Binary/halftone, gray-scale images
ISO JPEG	Still images
ISO JPEG2000	Digital cinema
ISO MPEG2	Digital video, SDTV, HDTV
ISO MPEG4 AVC/ITU-T H.264	Digital video
ISO HEVC/ ITU-T H.265	HD video, HDTV, UHDTV

2.4 3D Video

3D cinema has gained wide acceptance in theatres as many movies are now produced in 3D. Flat-panel 3DTV 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 3DTV 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

Stereoscopic (with glasses)	Auto-Stereoscopic (no glasses)	Lightfield (no glasses)	Volumetric (no glasses)
Color-multiplexed	Two-view	Super multi-view	Static volume
Polarization-multiplexed	Multi-view	Holoscopic (Integral)	Swept volume
Time-multiplexed	With head-tracking	Holographic	

Figure 2.10 Classification of 3D-display technologies.

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.

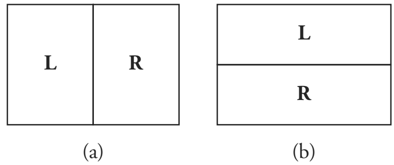

Figure 2.11 Frame compatible formats: (a) side-by-side; (b) top-bottom.

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%

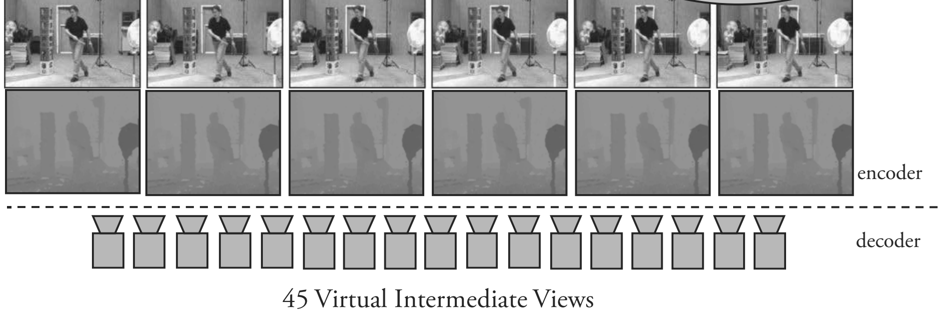

Figure 2.12 N-view + N depth-map format (courtesy of Aljoscha Smolic).

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 terrestial (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 Terrestial 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 terrestial 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 terrestial 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 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 handheld 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 manufactuers 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 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 [Gu 14]. 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 [Har 04, For 11, Sze 11]. 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 8×8 blocks. Coarse quantization of DC coefficients may cause mismatch of image mean over 8×8 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 higly 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 $\hat{s}[n_1, n_2, k]$, 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

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

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 $\mathbf{x}$ and $\mathbf{y}$, with means μ_x and μ_y and variances σ_x^2 and σ_y^2, respectively. It is given by [Wan 04]

$$SSIM(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

where σ_{xy} is the covariance between windows $\mathbf{x}$ and $\mathbf{y}$ 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

[Boe 13] Boev, A., R. Bregovic, and A. Gotchev, "Signal processing for stereoscopic and multiview 3D displays," chapter in *Handbook of Signal Processing Systems, Second Edition*, (ed. S. Bhattacharyya, E. Deprettere, R. Leupers, and J. Takala), pp. 3–47, New York, NY: Springer, 2013.

[Bov 13] Bovik, A. C., "Automatic prediction of perceptual image and video quality," Proc. of the IEEE, vol. 101, no. 9, pp. 2008–2024, Sept. 2013.

[Bre 07] Breitmeyer, B. G., "Visual masking: past accomplishments, present status, future developments," Adv. Cogn. Psychol 3, 2007.

[Cho 14] Choi, L. K., Y. Liao, A. C. Bovik, "Video QoE models for the compute continuum," IEEE ComSoc MMTC E-Letter, vol. 8, no. 5, pp. 26–29, Sept. 2013.

[Dub 10] Dubois, E., *The Structure and Properties of Color Spaces and the Representation of Color Images,* Morgan & Claypool, 2010.

[Fer 01] Ferwerda, J. A., "Elements of early vision for computer graphics, IEEE Computer Graphics and Application, vol. 21, no. 5, pp. 22–33, Sept./Oct. 2001.

[For 11] Forsyth, David A. and Jean Ponce, *Computer Vision: A Modern Approach, Second Edition*, Upper Saddle River, NJ: Prentice Hall, 2011.

[Gon 07] Gonzalez, Rafael C. and Richard E. Woods, *Digital Image Processing, Third Edition*, Upper Saddle River, NJ: Prentice Hall, 2007.

[Gra 10] Granados, M., B. Ajdin, M. Wand, C. Theobalt, H-P. Seidel, and H. P.A. Lensch, "Optimal HDR reconstruction with linear digital cameras," IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR), pp. 215–222, June 2010.

[Gu 14] Gu, Y., et al., Survey of P2P Streaming Applications, IETF draft-ietf-ppsp-survey-08, April 2014.

[Har 04] Hartley, R. I. and A. Zisserman, *Multiple View Geometry in Computer Vision, Second Edition*, New York, NY: Cambridge University Press, 2004.

[HDM 13] High definition multimedia interface (HDMI). http://www.hdmi.org/index.aspx

[Hub 88] Hubel, D. H., "Eye, Brain, and Vision, Vol. 22, Scientific American Library," distributed by W. H. Freeman & Co., New York, 1988. http://hubel.med.harvard.edu

[IEC 00] IEC 61966-2-1:2000, Multimedia systems and equipment - Colour measurement and management - Colour management - Default RGB colour space: sRGB, Sept. 2000.

[ITU 02] ITU-R Rec. BT.709, Parameter values for the HDTV standards for production and international program exchange, April 2002. http://www.itu.int/rec/R-REC-BT.709

[ITU 11] ITU-R Rec. BT.601, Studio encoding parameters of digital television for standard 4:3 and wide screen 16:9 aspect ratios, March 2011. http://www.itu.int/rec/R-REC-BT.601/

[ITU 12] ITU-R Rec. BT.2020, Parameter values for ultra-high definition television systems for production and international program exchange, August 2012. http://www.itu.int/rec/R-REC-BT.2020-0-201208-I

[Kau 07] Kauff, P., et al., "Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability," Signal Processing: Image Communication, vol. 22, pp. 217–234, 2007.

[Kov 14] Kovacs, P. T., A. Boev, R. Bregovic, and A. Gotchev, "Quality measurements of 3D light-field displays," Int. Workshop on Video Processing and Quality metrics for Consumer Electronics (VPQM), Chandler, AR, USA, Jan. 30–31, 2014.

[Led 13] Lederer, S., C. Muller, B. Rainer, C. Timmerer, and H. Hellwagner, "An experimental analysis of dynamic adaptive streaming over HTTP in content centric networks", in Proc. of IEEE Int. Conf. on Multimedia and Expo (ICME), San Jose, USA, July 2013.

[Mul 85] Mullen, K. T., "The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings," J. Physiol., 1985.

[Nes 67] van Nes, F. L., J. J. Koenderink, H. Nas, and M. A. Bouman, "Spatiotemporal modulation transfer in the human eye," Jour. of the Optical Society of America, vol. 57, no. 9, Sept. 1967.

[MPG 07] ISO/IEC 23002-3:2007 Information technology - MPEG video technologies - Part 3: Representation of auxiliary video and supplemental information, 2007.

[Mul 11] Muller, K., P. Merkle, and T. Wiegand, "3-D video representation using depth maps," Proc. of the IEEE, vol. 99, no. 4, pp. 643–656, April 2011.

[Pan 14] Pantos, R. and W. May, HTTP Live Streaming, IETF draft-pantos-http-live-streaming-14, October 2014.

[Pin 04] Pinson, M. and S. Wolf, "A new standardized method for objectively measuring video quality," IEEE Trans. on Broadcasting, vol. 50, no.3, pp. 312–322, Sept. 2004.

[Pit 13] Pitas, I., *Digital Video and Television,* Ioannis Pitas: 2013.

[Rei 07] Reinhard, E., T. Kunkel, Y. Marion, J. Brouillat, R. Cozot and K. Bouatouch, "Image display algorithms for high and low dynamic range display devices," Jour. of the Society for Information Display, vol. 15 (12), pp. 997–1014, 2007.

[See 04] Seetzen, H., W. Heidrich, W. Stuerzlinger, G. Ward, L. Whitehead, M. Trentacoste, A. Ghosh, and A. Vorozcovs, "High dynamic range display systems," Proc. ACM SIGGRAPH, 2004.

[Sha 13] Shao, F., W. Lin, S. Gu, G. Jiang, and T. Srikanthan, "Perceptual full-reference quality assessment of stereoscopic images by considering binocular visual characteristics." IEEE Trans. on Image Proc., vol. 22, no. 5, pp. 1940–53, May 2013.

[Sha 98] Sharma, G., M. J. Vrhel, and H. J. Trussell, "Color imaging for multimedia," Proc. of the IEEE, vol. 86, no. 6, June 1998.

[Smo 11] Smolic, A., "3D video and free viewpoint video - From capture to display," Pattern Recognition, vol. 44, no. 9, pp. 1958–1968, Sept. 2011.

[Suc 11] Suchow, J. W. and G. A. Alvarez, "Motion silences awareness of visual change," Curr. Biol., vol. 21, no. 2, pp.140–143, Jan. 2011.

[Sze 11] Szeliski, R., *Computer Vision: Algorithms and Applications*, New York, NY: Springer, 2011.

[Tru 93] Trussell, H. J., "DSP solutions run the gamut for color systems," IEEE Signal Processing Mag., pp. 8–23, Apr. 1993.

[Uka 08] Ukai, K., and P. A. Howarth, "Visual fatigue caused by viewing stereoscopic motion images: Background, theories, and observations," *Displays,* vol. 29, pp. 106–116, Mar. 2008.

[Ure 11] Urey, H., K. V. Chellapan, E. Erden, and P. Surman, "State of the art in stereoscopic and autostereoscopic displays," Proc. of the IEEE, vol. 99, no. 4, pp. 540–555, April 2011.

[Wad 96] Wade, N. J., "Descriptions of visual phenomena from Aristotle to Wheatstone," *Perception,* vol. 25, no. 10, pp. 1137–1175, 1996.

[Wan 95] Wandell, B., *Foundations of Vision,* Sunderland, MA: Sinauer Associates, 1995.

[Wan 04] Wang, Z., A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Trans. on Image Processing, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[Win 13] Winkler, S. and D. Min, "Stereo/multiview picture quality: Overview and recent advances," Signal Processing: Image Communication, vol. 28, no. 10, pp. 1358–1373, Nov. 2013.

This page intentionally left blank

Index

A

M