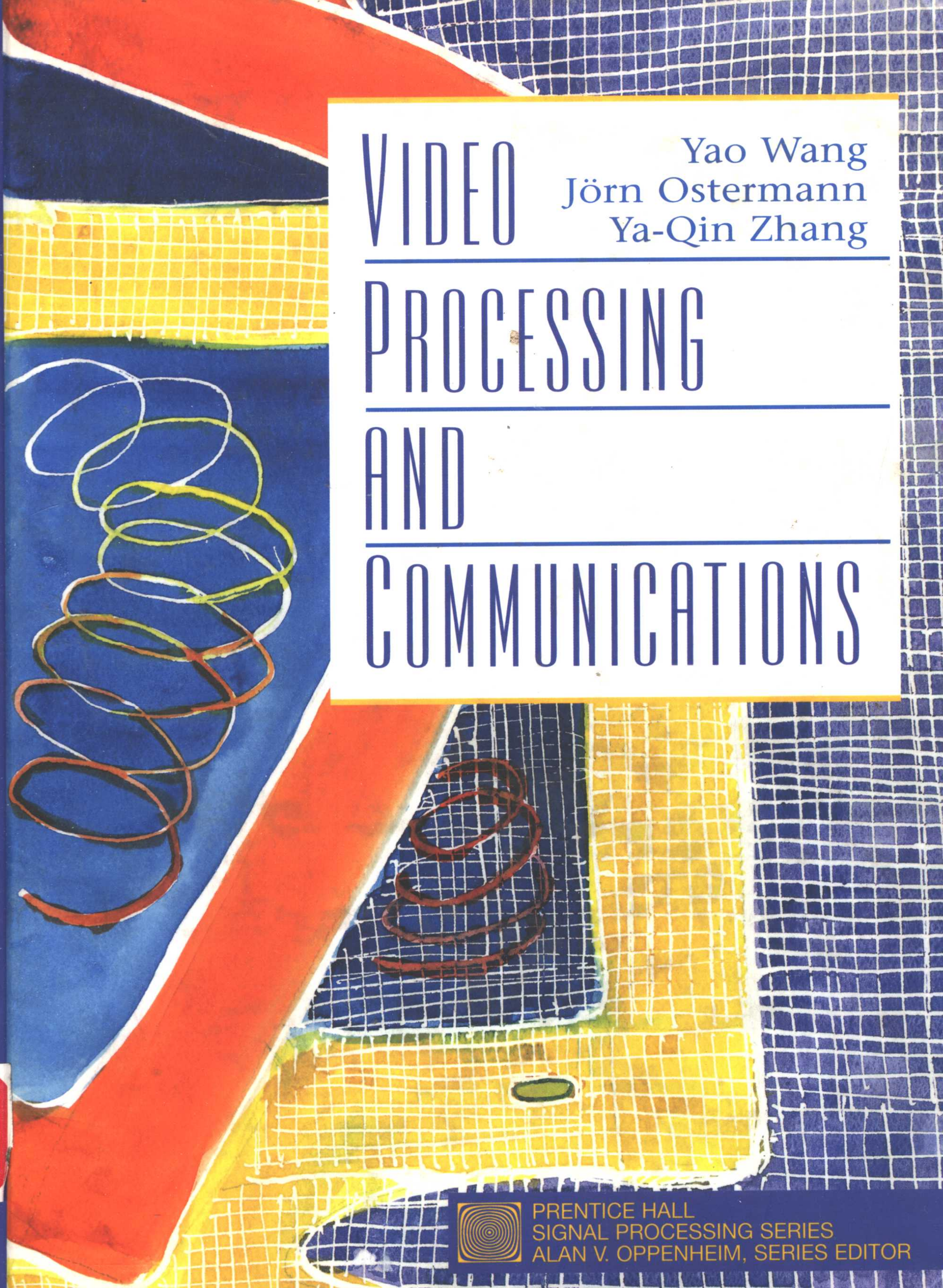
VIDEO PROCESSING AND COMMUNICATIONS

Yao Wang
Jörn Ostermann
Ya-Qin Zhang



PRENTICE HALL
SIGNAL PROCESSING SERIES
ALAN V. OPPENHEIM, SERIES EDITOR

VIDEO PROCESSING AND COMMUNICATIONS

Yao Wang

Jörn Ostermann

Ya-Qin Zhang

Prentice Hall

PRENTICE HALL

Upper Saddle River, New Jersey 07458

Library of Congress Cataloging-in-Publication Data

Wang, Yao, 1962–
Video processing and communications / Yao Wang, Jörn Ostermann, Ya-Qin Zhang.
p. cm.
Includes bibliographical references and index.
ISBN 0-13-017547-1
1. Image transmission. 2. Digital video I. Ostermann, Jörn. II. Zhang, Ya-Qin, 1969–
III. Title.

TK5105.2 .W36 2001
621.388'33—dc21

200103699

Vice President and Editorial Director, ECS: *Marcia J. Horton*
Publisher: *Tom Robbins*
Editorial Assistant: *Jessica Romeo*
Vice President and Director of Production and Manufacturing, ESM: *David W. Riccardi*
Executive Managing Editor: *Vince O'Brien*
Managing Editor: *David A. George*
Production Editor: *Kathy Ewing*
Director of Creative Services: *Paul Belfanti*
Creative Director: *Carole Anson*
Art Director: *Jayne Conte*
Art Editor: *Greg Dulles*
Manufacturing Manager: *Trudy Pisciotti*
Manufacturing Buyer: *Lynda Castillo*
Marketing Manager: *Holly Stark*
Marketing Assistant: *Karen Moon*

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

ISBN 0-13-017547-1

Pearson Education Ltd., *London*
Pearson Education Australia Pty. Ltd., *Sydney*
Pearson Education Singapore, Pte. Ltd.
Pearson Education North Asia Ltd., *Hong Kong*
Pearson Education Canada, Inc., *Toronto*
Pearson Educacíon de Mexico, S.A. de C.V.
Pearson Education—Japan, *Tokyo*
Pearson Education Malaysia, Pte. Ltd.

Preface

In the past decade or so, there have been fascinating developments in multimedia representation and communications. First of all, it has become very clear that all aspects of media are "going digital"; from representation to transmission, from processing to retrieval, from studio to home. Second, there have been significant advances in digital multimedia compression and communication algorithms, which make it possible to deliver high-quality video at relatively low bit rates in today's networks. Third, the advancement in VLSI technologies has enabled sophisticated software to be implemented in a cost-effective manner. Last but not least, the establishment of half a dozen international standards by ISO/MPEG and ITU-T laid the common groundwork for different vendors and content providers.

At the same time, the explosive growth in wireless and networking technology has profoundly changed the global communications infrastructure. It is the confluence of wireless, multimedia, and networking that will fundamentally change the way people conduct business and communicate with each other. The future computing and communications infrastructure will be empowered by virtually unlimited bandwidth, full connectivity, high mobility, and rich multimedia capability.

As multimedia becomes more pervasive, the boundaries between video, graphics, computer vision, multimedia database, and computer networking start to blur, making video processing an exciting field with input from many disciplines. Today, video processing lies at the core of multimedia. Among the many technologies involved, video coding and its standardization are definitely the key enablers of these developments. This book covers the fundamental theory and techniques for digital video processing, with a focus on video coding and communications. It is intended as a textbook for a graduate-level course on video processing, as well as a reference or self-study text for researchers and engineers. In selecting the topics to cover, we have tried to achieve a balance between providing a solid theoretical foundation and presenting complex system issues in real video systems.

SYNOPSIS

Chapter 1 gives a broad overview of video technology, from analog color TV system to digital video. Chapter 2 delineates the analytical framework for video analysis in the frequency domain, and describes characteristics of the human visual system. Chapters 3–12 focus on several very important sub-topics in digital video technology. Chapters 3 and 4 consider how a continuous-space video signal can be sampled to retain the maximum perceivable information within the affordable data rate, and how video can be converted from one format to another. Chapter 5 presents models for the various components involved in forming a video signal, including the camera, the illumination source, the imaged objects and the scene composition. Models for the three-dimensional (3-D) motions of the camera and objects, as well as their projections onto the two-dimensional (2-D) image plane, are discussed at length, because these models are the foundation for developing motion estimation algorithms, which are the subjects of Chapters 6 and 7. Chapter 6 focuses on 2-D motion estimation, which is a critical component in modern video coders. It is also a necessary preprocessing step for 3-D motion estimation. We provide both the fundamental principles governing 2-D motion estimation, and practical algorithms based on different 2-D motion representations. Chapter 7 considers 3-D motion estimation, which is required for various computer vision applications, and can also help improve the efficiency of video coding.

Chapters 8–11 are devoted to the subject of video coding. Chapter 8 introduces the fundamental theory and techniques for source coding, including information theory bounds for both lossless and lossy coding, binary encoding methods, and scalar and vector quantization. Chapter 9 focuses on waveform-based methods (including transform and predictive coding), and introduces the block-based hybrid coding framework, which is the core of all international video coding standards. Chapter 10 discusses content-dependent coding, which has the potential of achieving extremely high compression ratios by making use of knowledge of scene content. Chapter 11 presents scalable coding methods, which are well-suited for video streaming and broadcasting applications, where the intended recipients have varying network connections and computing powers. Chapter 12 introduces stereoscopic and multiview video processing techniques, including disparity estimation and coding of such sequences.

Chapters 13–15 cover system-level issues in video communications. Chapter 13 introduces the H.261, H.263, MPEG-1, MPEG-2, and MPEG-4 standards for video coding, comparing their intended applications and relative performance. These standards integrate many of the coding techniques discussed in Chapters 8–11. The MPEG-7 standard for multimedia content description is also briefly described. Chapter 14 reviews techniques for combating transmission errors in video communication systems, and also describes the requirements of different video applications, and the characteristics of various networks. As an example of a practical video communication system, we end the text with a chapter devoted to video streaming over the Internet and wireless network. Chapter 15 discusses the requirements and representative solutions for the major subcomponents of a streaming system.

SUGGESTED USE FOR INSTRUCTION AND SELF-STUDY

As prerequisites, students are assumed to have finished undergraduate courses in signals and systems, communications, probability, and preferably a course in image processing. For a one-semester course focusing on video coding and communications, we recommend covering the two beginning chapters, followed by video modeling (Chapter 5), 2-D motion estimation (Chapter 6), video coding (Chapters 8–11), standards (Chapter 13), error control (Chapter 14) and video streaming systems (Chapter 15). On the other hand, for a course on general video processing, the first nine chapters, including the introduction (Chapter 1), frequency domain analysis (Chapter 2), sampling and sampling rate conversion (Chapters 3 and 4), video modeling (Chapter 5), motion estimation (Chapters 6 and 7), and basic video coding techniques (Chapters 8 and 9), plus selected topics from Chapters 10–13 (content-dependent coding, scalable coding, stereo, and video coding standards) may be appropriate. In either case, Chapter 8 may be skipped or only briefly reviewed if the students have finished a prior course on source coding. Chapters 7 (3-D motion estimation), 10 (content-dependent coding), 11 (scalable coding), 12 (stereo), 14 (error-control), and 15 (video streaming) may also be left for an advanced course in video, after covering the other chapters in a first course in video. In all cases, sections denoted by asterisks (*) may be skipped or left for further exploration by advanced students.

Problems are provided at the end of Chapters 1–14 for self-study or as homework assignments for classroom use. Appendix D gives answers to selected problems. The website for this book (www.prenhall.com/wang) provides MATLAB scripts used to generate some of the plots in the figures. Instructors may modify these scripts to generate similar examples. The scripts may also help students to understand the underlying operations. Sample video sequences can be downloaded from the website, so that students can evaluate the performance of different algorithms on real sequences. Some compressed sequences using standard algorithms are also included, to enable instructors to demonstrate coding artifacts at different rates by different techniques.

ACKNOWLEDGMENTS

We are grateful to the many people who have helped to make this book a reality. Dr. Barry G. Haskell of AT&T Labs, with his tremendous experience in video coding standardization, reviewed Chapter 13 and gave valuable input to this chapter as well as other topics. Prof. David J. Goodman of Polytechnic University, a leading expert in wireless communications, provided valuable input to Section 14.2.2, part of which summarize characteristics of wireless networks. Prof. Antonio Ortega of the University of Southern

California and Dr. Anthony Vetro of Mitsubishi Electric Research Laboratories, then a Ph.D. student at Polytechnic University, suggested what topics to cover in the section on rate control, and reviewed Sections 9.3.3–4. Mr. Dapeng Wu, a Ph.D. student at Carnegie Mellon University, and Dr. Yiwei Hou from Fijitsu Labs helped to draft Chapter 15. Dr. Ru-Shang Wang of Nokia Research Center, Mr. Fatih Porikli of Mitsubishi Electric Research Laboratories, also a Ph.D. student at Polytechnic University, and Mr. Khalid Goudeaux, a student at Carnegie Mellon University, generated several images related to stereo. Mr. Haidi Gu, a student at Polytechnic University, provided the example image for scalable video coding. Mrs. Dorota Ostermann provided the brilliant design for the cover.

We would like to thank the anonymous reviewers who provided valuable comments and suggestions to enhance this work. We would also like to thank the students at Polytechnic University, who used draft versions of the text and pointed out many typographic errors and inconsistencies. Solutions included in Appendix D are based on their homeworks. Finally, we would like to acknowledge the encouragement and guidance of Tom Robbins at Prentice Hall. Yao Wang would like to acknowledge research grants from the National Science Foundation and New York State Center for Advanced Technology in Telecommunications over the past ten years, which have led to some of the research results included in this book.

Most of all, we are deeply indebted to our families, for allowing and even encouraging us to complete this project, which started more than four years ago and took away a significant amount of time we could otherwise have spent with them. The arrival of our new children Yana and Brandon caused a delay in the creation of the book but also provided an impetus to finish it. This book is a tribute to our families, for their love, affection, and support.

YAO WANG
Polytechnic University, Brooklyn, NY, USA
yao@vision.poly.edu

JÖRN OSTERMANN
AT&T Labs—Research, Middletown, NJ, USA
osterman@research.att.com

YA-QIN ZHANG
Microsoft Research, Beijing, China
yzhang@microsoft.com

Glossary of Notations

Mathematical notation	
$[\mathbf{A}]$	Boldface upper case letters in roman with square bracket represent matrices
$\mathbf{a}, \mathbf{A}$	Boldface lower or upper case letters in roman represent vectors
$[\mathbf{A}]^T, \mathbf{a}^T$	Transpose of matrices or vectors
$[\mathbf{A}]^{-1}$	Inverse of matrix $[\mathbf{A}]$
$\det[\mathbf{A}]$	Determinant of matrix $[\mathbf{A}]$
$\lVert\mathbf{a}\rVert$	l_2-norm of vector $\mathbf{a}$
$\mathcal{R}^K$	K-D real space
$\mathcal{C}^K$	K-D complex space
$\mathcal{Z}^K$	K-D integer space
$\mathcal{F}$	Upper case letters in script represent random variables
$\boldsymbol{\mathcal{F}}$	Boldface upper case letters in script represent random vectors, $\boldsymbol{\mathcal{F}} = [\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_N]^T$, or a random sequence, $\boldsymbol{\mathcal{F}} = \{\mathcal{F}_1, \mathcal{F}_2, \ldots, \}$
$E\{\cdot\}$	Expectation operation
$\forall$	For all
Λ	Representing a lattice
Λ^*	Reciprocal lattice of Λ
$d(\Lambda)$	Density of lattice Λ.
$[\mathbf{V}]$	Generating matrix for Λ
$[\mathbf{U}]$	Generating matrix for Λ^*, $[\mathbf{U}] = ([\mathbf{V}]^T)^{-1}$
$\mathcal{V}(\Lambda)$	Voronoi cell of Λ

Video representation	
$\mathbf{X} = [X, Y, Z]^T$	Coordinate of a point in the 3-D space: X, Y, and Z represent the horizontal, vertical, and depth positions, respectively
$\mathbf{x} = [x, y]^T$	Coordinate of a pixel in the 2-D image plane: x and y represent the horizontal and vertical positions, respectively
t	Time index, either continuous or discrete
$\mathbf{m} = [m, n]^T$	Discrete coordinate of a pixel in a 2-D digital image: m and n represent the column and row indices of pixels, respectively
k	Discrete time index (i.e., frame number)
$\psi(x, y, t)$	Image value at pixel location $\mathbf{x}$ and frame time t of a video sequence. The image value could be a scalar representing the luminance or a vector representing three color components. This notation is used to refer to a video signal in general, defined over either a continuous or discrete space
$\psi_1(x, y), \psi_2(x, y)$	Representing the anchor and target frames involved in 2-D motion estimation
$\psi_p(x, y, t)$	The predicted image of frame $\psi(x, y, t)$ in video coding
$\psi_r(x, y)$	The reference frame used to predict a frame
$\psi_v(x, y, t)$	Image function in view v in a multiview system
$\psi(m, n, k)$	Image value at pixel location (m, n) and frame time k in a digital video
$\nabla\psi$	Spatial gradient of $\psi(x, y, t)$, $\nabla\psi = [\frac{\partial\psi}{\partial x}, \frac{\partial\psi}{\partial y}]^T$
$\Delta_x, \Delta_y, \Delta_t$	Sampling interval in horizontal, vertical, and temporal directions, respectively
$f_{s,t}$	Sampling frequency in temporal direction or frame rate, $f_{s,t} = 1/\Delta_t$, measured in frames/s (fps) or Hz
$f_{s,x}, f_{s,y}$	Sampling frequencies in horizontal and vertical directions, $f_{s,x} = 1/\Delta_x$, $f_{s,y} = 1/\Delta_y$, commonly measured in pixels/picture-width and pixels/picture-height
R	Bit rate, specified in bits/s (bps) for a video sequence, bits/pixel (bpp) for an image, or bits/sample for a general discrete source

Frequency-domain representations	
f_x, f_y	Horizontal and vertical frequencies, usually measured in cycles/degree (cpd)
f_θ	Angular frequencies, measured in cpd
f_t	Temporal frequency, measured in cycles/s (cps) or Hz
$\mathbf{f}$	Frequency index in a multidimensional space. For video signals, $\mathbf{f} = [f_x, f_y, f_t]^t$
$\Psi_c(f_x, f_y, f_t)$	Continuous-space Fourier transform (CSFT) of $\psi(x, y, t)$
$\Psi_s(f_x, f_y, f_t)$	Sampled-space Fourier transform (SSFT) of $\psi(x, y, t)$
$\Psi_d(f_x, f_y, f_t)$	Discrete-space Fourier transform (DSFT) of $\psi(m, n, k)$
$\Psi(f_x, f_y, f_t)$	Generally used to refer to the Fourier transform of a video signal; can be CSFT, SSFT, or DSFT

Motion/disparity/camera characterization	
$\mathbf{D}(\mathbf{X}, t_1, t_2)$	3-D motion vector of point $\mathbf{X}$ from time t_1 to time t_2. When the underlying t_1 and t_2 are clear, we simply write $\mathbf{D}(\mathbf{X})$. The X-, Y-, and Z-components of $\mathbf{D}$ is denoted by D_x, D_y, D_z
$\mathbf{d}(\mathbf{x}, t_1, t_2)$	2-D motion vector of pixel $\mathbf{x}$ from time t_1 to time t_2. When the underlying t_1, t_2 are clear, we simply write $\mathbf{d}(\mathbf{x})$. The x– and y–components of $\mathbf{d}$ are denoted by d_x and d_y. $\mathbf{d}(\mathbf{x}; \mathbf{a})$ represents the motion field as a function of motion parameter vector $\mathbf{a}$. The same notation $\mathbf{d}(\mathbf{x})$ is also used to represent the disparity vector between two views in a stereo sequence
$\mathbf{w}(\mathbf{x})$	The mapping function between two image frames, $\mathbf{w}(\mathbf{x}) = \mathbf{x} + \mathbf{d}(\mathbf{x})$. $\mathbf{w}(\mathbf{x}; \mathbf{a})$ represents the mapping function as a function of motion parameters $\mathbf{a}$
$\theta_x, \theta_y, \theta_z$	Camera or object rotation angles with respect to X, Y, and Z axes, respectively, of a predefined world coordinate
$[\mathbf{R}]$	The rotation matrix of a camera or object in 3-D
$\mathbf{T}$	The translation vector of a camera or object in 3-D, $\mathbf{T} = [T_x, T_y, T_z]^T$
F	Camera focal length
$\mathbf{C}$	Camera center in the world coordinate

Contents