ANALYZING VIDEO SEQUENCES OF MULTIPLE HUMANS

Tracking, Posture Estimation and Behavior Recognition

THE KLUWER INTERNATIONAL SERIES IN VIDEO COMPUTING

Series Editor

Mubarak Shah, Ph.D.
University of Central Florida
Orlando, USA

Video is a very powerful and rapidly changing medium. The increasing availability of low cost, low power, highly accurate video imagery has resulted in the rapid growth of applications using this data. Video provides multiple temporal constraints, which make it easier to analyze a complex, and coordinated series of events that cannot be understood by just looking at only a single image or a few frames. The effective use of video requires understanding of video processing, video analysis, video synthesis, video retrieval, video compression and other related computing techniques.

The Video Computing book series will provide a forum for the dissemination of innovative research results for computer vision, image processing, database and computer graphics researchers, who are interested in different aspects of video.

ANALYZING VIDEO SEQUENCES OF MULTIPLE HUMANS

Tracking, Posture Estimation and Behavior Recognition

Jun Ohya
Waseda University

Akira Utsumi
Advanced Telecommunications Research Institute International

Junji Yamato
Nippon Telegraph & Telephone Corporation

SPRINGER SCIENCE+BUSINESS MEDIA, LLC

ISBN 978-1-4613-5346-1 ISBN 978-1-4615-1003-1 (eBook)
DOI 10.1007/978-1-4615-1003-1

Library of Congress Cataloging-in-Publication Data

A C.I.P. Catalogue record for this book is available
from the Library of Congress.

Originally published by Kluwer Academic Publishers in 2002
Softcover reprint of the hardcover 1st edition 2002

Printed on acid-free paper.

Contents

List of Figures

List of Tables

Foreword

Traditionally, scientific fields have defined boundaries, and scientists work on research problems within those boundaries. However, from time to time those boundaries get shifted or blurred to evolve new fields. For instance, the original goal of computer vision was to understand a single image of a scene, by identifying objects, their structure, and spatial arrangements. This has been referred to as *image understanding*. Recently, computer vision has gradually been making the transition away from understanding single images to analyzing image sequences, or *video understanding*. Video understanding deals with understanding of video sequences, e.g., recognition of gestures, activities, facial expressions, etc. The main *shift* in the classic paradigm has been from the recognition of static objects in the scene to motion-based recognition of actions and events. Video understanding has overlapping research problems with other fields, therefore *blurring* the fixed boundaries.

Computer graphics, image processing, and video databases have obvious overlap with computer vision. The main goal of computer graphics is to generate and animate realistic looking images, and videos. Researchers in computer graphics are increasingly employing techniques from computer vision to generate the synthetic imagery. A good example of this is image-based rendering and modeling techniques, in which geometry, appearance, and lighting is derived from real images using computer vision techniques. Here the *shift* is from *synthesis* to *analysis followed by synthesis*. Image processing has always overlapped with computer vision because they both inherently work directly with images. One view is to consider image processing as low-level computer vision, which *processes* images, and video for later analysis by high-level computer vision techniques. Databases have traditionally contained text, and numerical data. However, due to the current availability of video in digital form, more and more databases are containing video as content. Consequently, researchers in databases are increasingly applying computer vision techniques to analyze the video before indexing. This is essentially *analysis followed by indexing*.

Due to the emerging MPEG-4, and MPEG-7 standards, there is a further overlap in research for computer vision, computer graphics, image processing,

and databases. In a typical model-based coding for MPEG-4, video is first *analyzed* to estimate local and global motion then the video is *synthesized* using the estimated parameters. Based on the difference between the real video and synthesized video, the model parameters are *updated* and finally *coded* for transmission. This is essentially *analysis followed by synthesis, followed by model update, and followed by coding*. Thus, in order to solve research problems in the context of the MPEG-4 codec, researchers from different video computing fields will need to collaborate. Similarly, MPEG-7 will bring together researchers from databases, and computer vision to specify a standard set of descriptors that can be used to describe various types of multimedia information. Computer vision researchers need to develop techniques to automatically compute those descriptors from video, so that database researchers can use them for indexing.

Due to the overlap of these different areas, it is meaningful to treat *video computing* as one entity, which covers the parts of computer vision, computer graphics, image processing, and databases that are related to video. This international series on *Video Computing* will provide a forum for the dissemination of innovative research results in video computing, and will bring together a community of researchers, who are interested in several different aspects of video.

Mubarak Shah
University of Central Florida

Orlando
January 20, 2002

Preface

In recent years, **video** has become ubiquitous in daily life. The VCR has become one of the most widely used appliances, typically for recording television programs. Compact video cameras for home use are also very common. Although it has become easier to record video with such technologies, however, editing video can still be difficult or tedious for the average person (despite the development of editing software tools for the personal computer). As a result, VCR and video camera users may amass large quantities of raw, unedited footage that is seldom watched, because segments of interest to the user cannot be easily accessed. In another application, video cameras are also commonly used for surveillance of offices, shops, and homes, but these video streams may require continuous monitoring by security personnel, thus consuming valuable human resources and being prone to lapses of attention by the human observer. These examples suggest that there is a need for automatic analysis of the content of video footage (for example, to facilitate editing, retrieval, or monitoring). Such automation, ideally in real-time, would reduce the burden on the user and broaden the possible applications of video.

To pursue such goals, it is useful to take approaches from the field of **Computer Vision**, one of the most active areas of computer science, that develops algorithms to automatically analyze images acquired by cameras. Recent technical developments have enabled computer vision to deal with video sequences. Such computer vision based video analysis technologies will likely be utilized for a variety of applications, such as telecommunication, video compression, surveillance and security, advanced video games, indexing and retrieval of multimedia database systems, producing digital cinema, and editing video libraries.

This book focuses on **humans** as the subjects of video sequences. This focus is a natural consequence of the immensely important and meaningful role images of people play in daily life. Technically, video sequences of humans are a challenging target for computer vision algorithms, due to the following reasons.

- Multiple people can be in a scene at once.

- Each person may be moving.
- The human body is a 3D, non-rigid, deformable, articulated object. Therefore,
  - In a video segment a person may be in various postures, each with a dramatically different appearance.
  - Occlusions could occur. (e.g., when one body part hides another)
- Gesture and activities could vary each time they are performed, even if the subject intends to repeat the same gesture or activity.

To tackle these issues, many technologies in computer vision are needed, such as tracking, silhouette extraction, contour analysis, shape analysis, 3D reconstruction, posture estimation, and pattern recognition. Many researchers in computer vision have been attracted to these challenging problems and have been studying human image analysis. As a result of the effort made by some of the leading researchers in this area, the IEEE (Institute of Electrical and Electronics Engineers) International Conference on Face and Gesture Recognition is held every two years. This and other major computer vision conferences provide forums for presenting papers and holding discussions with researchers in this field. Thus, human image analysis is an active area in computer vision research.

It is impossible to describe all work relevant to human image analysis in this book. Therefore, we concentrate on multiple human tracking, body posture estimation, and behavior recognition. We hope that this book will be useful to our readers (some of whom may be considering or may have already undertaken related projects) and will accelerate the progress in the research areas we discuss.

JUN OHYA

Acknowledgements

First of all, the authors are very grateful that Prof. Mubarak Shah of the University of Central Florida, U.S.A. contacted Jun Ohya and encouraged him to write this book. Without Prof. Shah's proposal, suggestions, and encouragement, this book would not have been published.

The projects described in this book were developed through the efforts of many people, and were conducted in the following chronological order.

Work on behavior recognition using Hidden Markov Models (HMM) (Chapter 4) was started soon after Junji Yamato entered NTT (Nippon Telegraph & Telephone Corp., Japan) and joined J. Ohya's team in 1990. J. Yamato and J. Ohya belonged to the group led by Dr. Kenichiro Ishii, and they thank Dr. Ishii for his significant key idea of applying HMM to human behaviors in video sequences. J. Yamato thanks Shoji Kurakake (currently, at NTT DoCoMo), and Prof. Akira Tomono (currently, at Tokai University) for their contributions to the new Vector Quantization formulation. J. Yamato also thanks Dr. Hiroshi Murase of NTT for his collaboration in the application of this HMM based method to the content-based video database retrieval.

The posture estimation project (Chapter 3) was started soon after J. Ohya became the head of Department 1 of ATR (Advanced Telecommunication Research Institute) Media Integration & Communications Research Laboratories (ATR MIC), Kyoto, Japan, in 1996. Many people in Department 1 were engaged in this project. J. Ohya thanks Dr. Kazuyuki Ebihara (currently, at Japan Victor Corp.), Dr. Tatsumi Sakaguchi (currently, at Sony Corp., Japan), Prof. Jun Kurumisawa (currently, at Chiba University of Commerce, Japan), Dr. Shoichiro Iwasawa (currently, at the Telecommunications Advancement Organization of Japan), Prof. Kazuhiko Takahashi (currently, at Yamaguchi University, Japan), Masanori Yamada (currently, at NTT), Katsuhiro Takematsu (currently, at Sony Corp.) and Tetsuya Uemura (currently, at Sony Corp.) for their contributions to this project.

The human tracking project (Chapter 2) was conducted mainly by Akira Utsumi, who was a member of J. Ohya's department at ATR MIC. A. Utsumi and J. Ohya thank some student interns: Hiroki Mori (currently, at NTT), Yang Howard (currently, at University of British Columbia, Canada), Mami Kinoshita (currently, at Nagaoka University of Technology, Japan), and Hirotake Yamazoe (from Osaka University, Japan). A. Utsumi thanks Dr. Nobuji Tetsutani, the current head of Department 1 of ATR MIC for his support of this project.

J. Ohya and A. Utsumi appreciate the efforts of Dr. Ryohei Nakatsu, the director of ATR MIC, who supervised the projects described in Chapters 2 and 3. J. Ohya and A. Utsumi thank CSK Corp.'s programmers: Hiroshi Urainkyo,

Shigeo Imura and Yuji Fujimoto (currently, at Sony Corp.) for their excellent coding.

The authors would like to express their gratitude to Kluwer Academic Publishers, who allowed them to write this book. J. Ohya thanks Michael Kowalski at Brown University, U.S.A. (formerly, in J. Ohya's department at ATR MIC), who kindly edited the manuscripts for Preface, Chapter 1, Chapter 3, and Chapter 5. Finally, the authors appreciate their families' patience and cooperation during the time period in which they were intensively writing this book.

JUN OHYA, AKIRA UTSUMI, JUNJI YAMATO

Contributing Authors

Jun Ohya is a professor at the Global Information and Telecommunication Institute, Waseda University, Tokyo Japan. He got his B.S., M.S., and Ph.D. degrees in Precision Machinery Engineering from the University of Tokyo in 1977, 1979, and 1988, respectively. In 1979, he entered NTT (Nippon Telegraph & Telephone Corp) Telecommunication Laboratories and worked on full color printing technologies and computer vision related technologies. From 1988 to 1989, he was a visiting research associate of the Computer Vision Laboratory, University of Maryland, College Park, Maryland, USA. In 1992, he transferred to ATR (Advanced Telecommunications Research Institute International) Communication Systems Research Laboratories, Kyoto, Japan and worked on combining virtual reality technologies with video conferencing systems. From 1996 to 2000, he was a department head of ATR Media Integration & Communications Research Laboratories, Kyoto, Japan, and supervised research projects on computer vision, virtual reality, and computer graphics, as well as integrating art and technologies. In 2000, he joined Waseda University. His current research interest includes multimedia science based on computer vision, computer graphics, virtual reality and integration of art and technologies.

Akira Utsumi received his B.E. degree in Matallurgical Engineering from Osaka Prefecture University in 1991. He received his M.E. degree in Information & Computer Siences and his Ph.D. degree (Engineering) from Osaka University in 1993 and1999, respectively. In 1993, he joined the Communication Systems Research Laboratories at Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan. From 1995 to 2001, he was with ATR Media Integration & Communications Research Laboratories. He is now a resercher at Media Information Science Laboratories at ATR. His current research interests include Computer Vision, Image Processing and Human-Computer Interaction.

Junji Yamato received B.Eng and M.Eng degrees in precision machinary engineering from the Univerisity of Tokyo, Japan, in 1988 and 1990, respectively. He received a M.S. degree in electrical engineering and computer science from the Massachusetts Institute of Technology in 1998, and a Ph.D. degree from the University of Tokyo in 2001. He was with the NTT Human Interface Laboratories, from 1990 to 1996, the MIT Artificial Intelligence Laboratory from 1996 to 1998, and the NTT Communication Science Laboratories from 1998 to 2001. He is currently a manager at the R &D strategy department of NTT Corp. His research interests include computer vision, machine learning, and human-robot interaction.