FACE DETECTION AND GESTURE RECOGNITION FOR HUMAN-COMPUTER INTERACTION

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 provides 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.

FACE DETECTION AND GESTURE RECOGNITION FOR HUMAN-COMPUTER INTERACTION

by

Ming-Hsuan Yang
Honda R&D Americas, Inc.
Honda Fundamental Research Laboratories

Narendra Ahuja
Beckman Institute and Department of Computer Science
University of Illinois at Urbana-Champaign

SPRINGER SCIENCE+BUSINESS MEDIA, LLC

Library of Congress Cataloging-in-Publication Data

Yang, Ming-Hsuan.
Face detection and gesture recognition for human-computer interaction / by Ming-Hsuan Yang, Narendra Ahuja.
p. cm—(The Kluwer international series in video computing ; 1)
Includes bibliographical references and index.
ISBN 978-1-4613-5546-5 ISBN 978-1-4615-1423-7 (eBook)
DOI 10.1007/978-1-4615-1423-7
1. Human-computer interaction. 2. Image processing—Digital techniques. I. Ahuja, Narendra, 1950- II. Title. III. Series.

QA76.9.H85 Y36 2001
004'.01'9—dc21

2001033814

Originally published by Kluwer Academic Publishers in 2001
Softcover reprint of the hardcover 1st edition 2001

Printed on acid-free paper.

Series 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
March 26, 2001

Preface

With the ubiquity of new information technology and media, more effective and friendly methods for human computer interaction (HCI) are being developed which do not rely on traditional devices such as keyboards, mice, and displays. Furthermore, the ever decreasing price/performance ratio of computing coupled with recent decreases in video image acquisition cost imply that computer vision systems can be deployed in desktop and embedded systems. The rapidly expanding research in face processing is based on the premise that information about a user's identity, state, and intent can be extracted from images, and that computers can then react accordingly, e.g., by observing a person's facial expression.

In the last few years, face detection/recognition as well as hand gesture recognition have attracted much attention though they have been studied for more than twenty years by psychophysicists, neuroscientists, and engineers. Many research demonstrations and commercial applications have been developed from these efforts.

A first step in any face processing system is detecting the locations in images where faces are present. However, face detection from a single image is a challenging task because of variability in scale, location, orientation (up-right, rotated), and pose (frontal, profile). Facial expression, occlusion, and lighting conditions also change the overall appearance of faces. Evidently, face detection is the first step in any automated system which solves the above problems.

Since Johansson's seminal work that suggests human movements can be recognized solely by motion information, motion profiles and trajectories have been investigated to recognize human motion by several researchers. Among the existing vision-based man-machine interfaces, hand gesture provides a natural and easy way for humans to command computers. Towards this, many methods have been developed to recognize the complex motion in hand gestures.

Many interesting and promising methods have been developed although all of these areas are still subject of active research. This book presents the work

on face detection done by the authors as well as summarizes existing other work in these areas. We first review work on the works on face detection in Chapter 2 and present some future research directions. We then present our algorithm for extracting and recognizing motion patterns in Chapter 3. Experimental results on extraction and recognition of motion patterns associated with hand gestures using a set of 40 ASL signs are presented. In Chapter 4, we describe a skin color model using a mixture of Gaussians. This model facilitates extraction of skin-tone regions, thereby reducing the computation in gesture recognition and face detection. Two algorithms using multimodal density functions are introduced in Chapter 5 to detect faces in gray-scale images. Experimental results on several benchmark databases are presented. In Chapter 6, we describe a SNoW-based face detector and explain why and when SNoW performs well. We conclude in Chapter 7 with observations about possible role of face detection and gesture recognition in intelligent human computer interaction.

MING-HSUAN YANG AND NARENDRA AHUJA

Acknowledgments

We would like to express our gratitude to numerous people who, in one way or another, have helped with the process of writing this book. Particularly, we would like to thank the following people for reviewing draft material for this book, and for discussions which have influenced parts of the text: David Kriegman, Tom Huang, Dan Roth and Mubarak Shah. Finally, we would like to thank staff at Kluwer Academic Press for their help in the final stages of preparing this book.

Contents