Thomas B. Moeslund
Adrian Hilton
Volker Krüger
Leonid Sigal *Editors*

Visual Analysis of Humans

Looking at People

Springer

Visual Analysis of Humans

Thomas B. Moeslund • Adrian Hilton •
Volker Krüger • Leonid Sigal
Editors

Visual Analysis of Humans

Looking at People

Springer

Editors
Assoc. Prof. Thomas B. Moeslund
Department of Media Technology
Aalborg University
Niels Jernes Vej 14
Aalborg, 9220
Denmark
tbm@create.aau.dk

Assoc. Prof. Volker Krüger
Copenhagen Institute of Technology
Aalborg University
Lautrupvang 2B
Ballerup, 2750
Denmark
vok@cvmi.aau.dk

Prof. Adrian Hilton
Centre for Vision, Speech & Signal Proc.
University of Surrey
Guildford, Surrey, GU2 7XH
UK
a.hilton@surrey.ac.uk

Dr. Leonid Sigal
Disney Research
Forbes Avenue 615
Pittsburgh, PA 15213
USA
lsigal@disneyresearch.com

ISBN 978-0-85729-996-3 e-ISBN 978-0-85729-997-0
DOI 10.1007/978-0-85729-997-0
Springer London Dordrecht Heidelberg New York

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2011939266

Cover design: VTeX UAB, Lithuania

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

Understanding human activity from video is one of the central problems in the field of computer vision. It is driven by a wide variety of applications in communications, entertainment, security, commerce and athletics. At its foundations are a set of fundamental computer vision problems that have largely driven the great progress that the field has made during the past few decades. In this book, the editors have assembled many of the world's leading authorities on video analysis of humans to assemble a comprehensive and authoritative set of chapters that cover both the core computer vision problems and the wide range of applications that solutions to these problems would enable.

The book is divided in four parts that cover detection and tracking of humans in video, measure human pose and movement from video, using these measurements to infer the activities that people are participating in, and finally describing the main applications areas that are based on these technologies. The book would be an excellent choice for a second graduate course on computer vision, or for a seminar on video analysis of human movement and activities. The combination of chapters that survey fundamental problems with others that go deeply into current approaches to topics provides the book with the excellent balance needed to support a well balanced course.

Part I, edited by Thomas B. Moeslund, focuses on problems associated with detecting and tracking people through camera networks. The chapter by Al Haj et al. discusses how multiple cameras can be cooperatively controlled so that people can both be tracked over large areas with cameras having wide fields of view and simultaneously imaged at high enough resolution with other cameras to analyze their activities. The next two chapters discuss two different approaches to detecting people in video. The chapter by Elgammal discusses background subtraction. Most simply, for a stationary camera one can detect moving objects by first building a model (an image) of an empty scene and then differencing that model with incoming video frames. Where the differences are high, movement has occurred. In reality, of course, things are much more complicated since the background can change over different time scales(due to wind load on vegetation, bodies of water in the scene, or the introduction of a new object into the background), the camera might

be active (panning, for example, as in Chap. 1), and there are many nuisance variables like shadows and specular reflections that should be eliminated. And, even if background subtraction worked "perfectly" true scene motion can be due to not just human movements but movement of other object in the scene. The chapter by Leibe, then, discusses a more direct method for detecting humans based on matching models of what people look like directly to the images. These methods generally employ a sliding window algorithm in which features based on shape and texture are combined to construct a local representation that can be used by statistical inference models to perform detection. Such methods can be used even when the camera is moving in an unconstrained manner. Detecting people is especially challenging because of variations in body shape, clothing and posture. The chapter by Chellappa discusses method for face detection. It not only contains an excellent introduction to sliding window based methods for face detection, but also explains how contextual information and high level reasoning can be used to locate faces, augmenting purely local approaches. The chapter by Song et al. discusses tracking. Tracking is especially complicated in situations where the scene contains many moving people because there is inevitably inter-occlusion. The chapter discusses fundamental multi-object techniques based on particle filters and joint probabilistic data association filters, and then goes on to discuss tracking in camera networks. Finally the chapter by Ellis and Ferryman discusses the various datasets that have been collected that researchers regularly use to evaluate new algorithms for detection and tracking.

Part II, edited by Leonid Sigal, discusses problems related to determining the time-varying 3D pose of a person from video. These problems have been intensely investigated over the past fifteen years and enormous progress has been made on designing effective and efficient representations for human kinematics, shape representations that can be used to model a wide variety of human forms, expressive and compact mathematical modeling mechanisms for natural human motion that can be used both for tracking and activity recognition, and computationally efficient algorithms that can be used to solve the nonlinear optimization problems that arise in human pose estimation and tracking. The first chapter by Pons-Mill and Rosenhahn begins by introducing criteria that characterize the utility of a parameterization of human pose and motion, and then discusses the merits of alternative representations with respect to these criteria. This is concerned with the "skeletal" component of the human model, and the chapter then goes on to discuss approaches to modeling the shapes of body parts. Finally, they discuss particle tracking methods that from an initial estimate of pose in a video sequence can both improve that estimate and then track the pose through the sequence. This process is illustrated for the case where the person can be segmented from the background, so that the silhouette of the person in each frame is (approximately) available. In the second chapter, Fleet motivates and discusses the use of low-dimensional latent models in pose estimation and tracking. While the space of all physically achievable human poses and motions might be very large, the poses associated with typical activities like walking lie on much lower-dimensional manifolds. The challenge is to identify representations that can simultaneously and smoothly map many activities to low-dimensional pose and

motion manifolds. Fleet's chapter discusses the Gaussian Process Latent Variable Model, along with a number of extensions to that model, that address this challenge. He also discusses the use of physics based models that, at least for well studied movements like walking, can be used to directly construct motion models to control tracking rather than learn them from large databases of examples. Ramanan provides an excellent introduction to parts based graphical models and methods to efficiently learn and solve for those models in images that can handle occlusion and appearance symmetries. Parts based models are especially relevant in situations where prior segmentation of a person from the background is not feasible—for example, for a video taken from a moving camera. They have been successfully applied to many object recognition problems; The chapter by Sminchisescu discusses methods that directly estimate (multi-valued) pose estimates from image measurements. Unlike the methods in the previous chapter that require complex search through the space of poses and motions, the methods here construct a direct mapping from images to poses (and motions). The main drawback of these algorithms is their limited ability to generalize to poses and motions not adequately represented in their training datasets. The approach described is a very general structure learning approach which is applicable to a wide variety of problems in computer vision. Finally, the chapter by Andriluka and Black discusses datasets for pose estimation and tracking as well as the criteria typically used by researchers to compare and evaluate algorithms.

Part III, edited by Volker Krüger, deals with the problems of representation and recognition of human (and vehicular) actions. For highly stylized or constrained actions (gestures, walking) one can approach the problem of recognizing them using, essentially, the same representations and recognition algorithms employed for static object detection and recognition. So, researchers have studied action recognition representations based on space time tubes of flow, or shape information captured by gradient histograms, or collections of local features such as 3D versions of SIFT, or "corners" on the 3D volume swept out by a dynamic human silhouette. All of these representations attempt to implicitly capture changing pose properties; however, for many actions it is sufficient to represent only the changing location of the person without regard to articulation—for example, to decide if one person is following another or if two people are approaching each other. The chapters by Wang, Nayak and Chowdhury contain complementary discussions of representations that can be used directly for appearance based action recognition. While Wang focuses on topic models as an inference model for activity recognition, Nayak et al. and Chowdhury contain surveys of other methods that have been frequently employed to represent, learn and recognize action classes, such as Hidden Markov Models or stochastic context free grammars. These more structured models are based on a decomposition of observations into motion "primitives" and the chapter by Kulic et al. discusses how these primitives might be represented and learned from examples. The chapter by Chowdhury also discusses the important problem of anomaly detection—finding instances of activities that are, in some way, performed differently from the norm. The problem of anomaly representation and detection is critical in many surveillance and safety applications (is a vehicle being driven erratically? has a pot been left on a stove too long?) and is starting to receive considerable attention in the

computer vision field. The chapter by Kjellström addresses the important problem of how context can be used to simultaneously improve action and object recognition. Many objects, especially at low magnification, look similar—consider roughly cylindrical objects like drinking glasses, flashlights, power screwdrivers. They are very hard to distinguish from one another based solely on appearance; but they are used in very different ways, so ambiguity about object class can be reduced through recognition of movements associated with human interaction with an object. Symmetrically, the body movements associated with many actions looks similar, but can be more easily differentiated by recognizing the objects that are used to perform the action. Kjellström explains how this co-dependence can be represented and used to construct more accurate vision systems. De la Torre's chapter covers the problem of facial expression recognition. Scientists have been interested in the problem of how and whether facial expressions reveal human internal state for over 150 years dating back to seminal work by Duchenne and Darwin on the subject in the 19th century. Paul Ekman's Facial Action Coding System (FACS) is an influential system to taxonimize people's facial expressions that has proven very useful to psychologists to model human behavior. Within the computer vision there has been intensive efforts to recognize and measure human facial expressions based on FACS and other models, and this chapter provides a comprehensive overview of the subject. Finally, Liu et al. discuss datasets that have been collected to benchmark algorithms for human activity recognition.

Finally, Part IV, edited by Adrian Hilton, contains articles describing some of the most important applications of activity recognition. The chapter by Chellappa is concerned with biometrics and discusses challenges and basic technical approaches to problems including face recognition, iris recognition and person recognition from gait. Human activity analysis to central to the design of monitoring systems for security and safety. Gong et al. discuss a variety of applications in surveillance including intruder detection, monitor public spaces for safety violations (such as left bag detection), and crowd monitoring (to different between normal crowd behavior and potentially disruptive behavior). Many of these applications depend on the ability of the surveillance system to accurately track individual people in crowded conditions for extended periods. Pellegrini's chapter discusses how simulation based motion models of human walking behavior in moderately crowded situations can be used to improve tracking of individuals. This is a relatively new area of research, and while current methods do not provide significant improvements in tracking accuracy over more classical methods, this is still an area with good potential to substantially improve tracking performance. Face and hand or body gesture recognition can be used to build systems that allow people to control computer applications in novel and natural ways, and Lin's chapter discusses fundamental methods for representing and recognizing face gestures (gaze, head pose) and hand gestures. Pantic's chapter addresses the interpretation of facial and body gestures in the context of human interactions with one another and their environment. They describe the exciting new research area of social signal processing—for example, determining whether participants in a discussion are agreeing or disagreeing, if there is a natural leader, or natural subgroups. The chapter provides a stimulating discussion of the basic

research problems and methodological issues in this emerging area. Another important application of face and gesture recognition is recognition of sign language, and the chapter by Cooper et al. contains an excellent introduction to this subject. While specialized devices like 3D data gloves can be used as input for hand sign language recognition systems, in typical situations where such devices are not available one has to address technically challenging problems of measuring hand geometry and motion from video. Additionally, sign languages are multi-modal—for example, they might include in addition to hand shape, arm motions and facial expressions. The chapter also discusses the research problems associated with developing systems for multi-modal sign recognition. Thomas's chapter discusses application in sports. Many applications require that players be tracked through the game—for example, to forensically determine how players react under different game conditions for strategy planning. Typically, these multi-agent tracking problems are addressed using multiple camera systems and one important practical problem that arises is controlling and calibrating these systems. The chapter by Grau discusses these multi-perspective vision problems in detail. Other applications require detailed tracking of a player's posture during play—for example to identify inefficiencies in a pitcher's throwing motions. There are many important applications of face and gesture analysis in the automotive industry—for example, determining the level of awareness of a driver, or where her attention is focused. The chapter by Tran and Trevedi summarizes the many ways that computer vision can be used to enhance driving safety and the approaches that researchers have employed to develop driver monitoring systems.

In summary, this is a timely collection of scholarly articles that simultaneously surveys foundations of human movement representation and analysis, illustrates these foundations in a variety of important applications and identifies many areas for fertile future research and development. The editors are to have congratulations on the exceptional job they did in organizing this volume.

College Park, USA Larry Davis
May 2011

Preface

Over the course of the last 10–20 years the field of computer vision has been preoccupied with the problem of looking at people. Hundreds, if not thousands, of papers have been published on the subject that span people and face detection, pose estimation, tracking and activity recognition. This research focus has been motivated by the numerous potential application for visual analysis of people from human–computer interaction to security, assisted living and clinical analysis of movement. A number of specific and general surveys have been published on these topics, but the field is lacking one coherent text that introduces and gives a comprehensive review of progress and open-problems. To provide such an overview is the exact ambition of this book. The target audience is not only graduate students in the computer vision field, but also scholars, researchers and practitioners from other fields who have an interest in systems for visual analysis of humans and corresponding applications.

The book is a collection of chapters that are written specifically for this book by leading experts in the field. Chapters are organized into four parts.

Part I: Detection and Tracking (seven chapters),
Part II: Pose Estimation (six chapters),
Part III: Recognition of Action (seven chapters),
Part IV: Applications (ten chapters).

The first three parts focus on different methods and the last part presents a number of different applications. The first chapter in each book part is an introduction chapter setting the scene. To support the reading of the book an index and list of glossary terms can be found in the back of the book. We hope this guide to research on the visual analysis of people contributes to future progress in the field and successful commercial application as the science and technology advances.

The editors would like to thank the authors for the massive work they have put into the different chapters! Furthermore we would like to thank Simon Rees and Wayne Wheeler from Springer for valuable guidance during the entire process of putting this book together. And finally, we would like to thank the reviewers who have helped to ensure the high standard of this book: Saiad Ali, Tamim Asfour, Patrick Buehler, Bhaskar Chakraborty, Rama Chellappa, Amit K.

Roy Chowdhury, Helen Cooper, Frederic Devernay, Mert Dikmen, David Fleet, Andrew Gilbert, Shaogang Gong, Jordi Gonzàlez, Jean-Yves Guillemaut, Abhinav Gupta, Ivan Huerta, Joe Kilner, Hedvig Kjellström, Dana Kulic, Bastian Leibe, Haowei Liu, Sebastien Marcel, Steve Maybank, Vittorio Murino, Kamal Nasrollahi, Eng-Jon Ong, Maja Pantic, Vishal Patel, Nick Pears, Norman Poh, Bodo Rosenhahn, Imran Saleemi, Mubarak Shah, Cristian Sminchisescu, Josephine Sullivan, Tai-Peng Tian, Sergio Valastin, Liang Wang, David Windridge, Ming-Hsuan Yang.

Aalborg University, Denmark — Thomas B. Moeslund
University of Surrey, UK — Adrian Hilton
Aalborg University, Denmark — Volker Krüger
Disney Research, Pittsburgh, USA — Leonid Sigal
May 2011

Contents

Contributors

Mykhaylo Andriluka Max Planck Institute for Computer Science, Saarbrücken, Germany, andriluka@mpi-inf.mpg.de

Michael J. Black Max Planck Institute for Intelligent Systems, Tübingen, Germany, black@tuebingen.mpg.de; Department of Computer Science, Brown University, Providence, USA, black@cs.brown.edu

Liefeng Bo University of Washington, Seattle, USA, lfb@cs.washington.edu

Aaron Bobick Georgia Institute of Technology, Atlanta, GA, USA, afb@cc.gatech.edu

Richard Bowden University of Surrey, Guildford, GU2 7XH, UK, R.Bowden@surrey.ac.uk

Rama Chellappa Department of Electrical and Computer Engineering, and UMIACS, University of Maryland, College Park, MD 20742, USA, rama@umiacs.umd.edu

Jeffrey F. Cohn Department of Psychology, University of Pittsburgh, Pittsburgh, PA 15260, USA, jeffcohn@pitt.edu

Helen Cooper University of Surrey, Guildford, GU2 7XH, UK, H.M.Cooper@surrey.ac.uk

Roderick Cowie Psychology Dept., Queen University Belfast, Belfast, UK

Francesca D'Errico Dept. Of Education, University Roma Tre, Rome, Italy

Larry Davis College Park, USA

Ahmed Elgammal Rutgers University, New Brunswick, NJ, USA, elgammal@cs.rutgers.edu

Anna-Louise Ellis University of Reading, Whiteknights, Reading, UK, a.l.ellis@reading.ac.uk