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

© Springer-Verlag London Limited 2011
Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms of licenses issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.
The use of registered names, trademarks, etc., in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant laws and regulations and therefore free for general use.
The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

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

Andreas Ess ETH Zürich, Zürich, Switzerland, aess@vision.ee.ethz.ch

Rogerio Feris IBM T.J. Watson Research Center, Hawthorn, NY 10532, USA, rsferis@ibm.com

Carles Fernández Computer Vision Center, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain, perno@cvc.uab.es

James Ferryman University of Reading, Whiteknights, Reading, UK, j.m.ferryman@reading.ac.uk

David J. Fleet Department of Computer Science, University of Toronto, Toronto, Canada, fleet@cs.toronto.edu

Shaogang Gong Queen Mary University of London, London, E1 4NS, UK, sgg@eecs.qmul.ac.uk

Jordi Gonzàlez Computer Vision Center and Departament de Ciències de la Computació, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain, jordi.gonzalez@uab.cat

Luc Van Gool ETH Zürich, Zürich, Switzerland, vangool@vision.ee.ethz.ch; KU Leuven, Leuven, Belgium

Raghuraman Gopalan Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742, USA, raghuram@umiacs.umd.edu

Oliver Grau BBC Research & Development, 56 Wood Lane, London, UK, Oliver.Grau@bbc.co.uk

Murad Al Haj Computer Vision Center, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain, malhaj@cvc.uab.es

Dirk Heylen EEMCS, University of Twente, Enschede, The Netherlands

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

Brian Holt University of Surrey, Guildford, GU2 7XH, UK, B.Holt@surrey.ac.uk

Thomas Huang Beckman Institute, University of Illinois at Urbana-Champaign, 405 North Mathews Avenue, Urbana, IL 61801, USA, huang@ifp.uiuc.edu

Ivan Huerta Computer Vision Center, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain, ivan.huerta@cvc.uab.es

Catalin Ionescu INS, University of Bonn, Bonn, Germany, catalin.ionescu@ins.uni-bonn.de

Atul Kanaujia ObjectVideo, Reston, VA, USA, atul.kanaujia@objectvideo.com

Hedvig Kjellström (Sidenbladh) CSC/CVAP, KTH, SE-100 44 Stockholm, Sweden, hedvig@kth.se

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

Danica Kragic Centre for Autonomous Systems, Royal Institute of Technology – KTH, Stockholm, Sweden, dani@kth.se

Dana Kulić University of Waterloo, Waterloo, Canada, dkulic@ece.uwaterloo.ca

Vuong Le Beckman Institute, University of Illinois at Urbana-Champaign, 405 North Mathews Avenue, Urbana, IL 61801, USA, vuongle2@ifp.uiuc.edu

Bastian Leibe UMIC Research Centre, RWTH Aachen University, Aachen, Germany, leibe@umic.rwth-aachen.de

Dennis Lin Beckman Institute, University of Illinois at Urbana-Champaign, 405 North Mathews Avenue, Urbana, IL 61801, USA, djlin@ifp.uiuc.edu

Haowei Liu University of Washington, Seattle, WA 98195, USA, hwliu@uw.edu

Chen Change Loy Queen Mary University of London, London, E1 4NS, UK, ccloy@eecs.qmul.ac.uk

Marc Mehu Psychology Dept., University of Geneva, Geneva, Switzerland

Thomas B. Moeslund Department of Architecture, Design and Media Technology, Aalborg University, Niels Jernes Vej 14, Aalborg, 9220, Denmark, tbm@create.aau.dk

Nandita M. Nayak University of California, Riverside, 900 University Ave. Riverside, CA 92521, USA, nandita.nayak@email.ucr.edu

Maja Pantic Computing Dept., Imperial College London, London , UK, m.pantic@imperial.ac.uk; EEMCS, University of Twente, Enschede, The Netherlands

Vishal M. Patel Department of Electrical and Computer Engineering, Center for Automation Research, University of Maryland, College Park, MD 20742, USA, pvishalm@umiacs.umd.edu

Catherine Pelachaud CNRS, Paris, France

Stefano Pellegrini ETH Zürich, Zürich, Switzerland, stefpell@vision.ee.ethz.ch

Jaishanker K. Pillai Department of Electrical and Computer Engineering, Center for Automation Research, University of Maryland, College Park, MD 20742, USA, jsp@umiacs.umd.edu

Isabella Poggi Dept. Of Education, University Roma Tre, Rome, Italy

Gerard Pons-Moll Leibniz University, Hanover, Germany, pons@tnt.uni-hannover.de

Deva Ramanan Department of Computer Science, University of California, Irvine, USA, dramanan@ics.uci.edu

Xavier Roca Computer Vision Center and Departament de Ciències de la Computació, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain, xavier.roca@uab.cat

Bodo Rosenhahn Leibniz University, Hanover, Germany, rosenhahn@tnt.uni-hannover.de

Amit K. Roy-Chowdhury University of California, Riverside, 900 University Ave. Riverside, CA 92521, USA, amitrc@ee.ucr.edu

Marc Schroeder DFKI, Saarbrucken, Germany

William R. Schwartz Institute of Computing, University of Campinas, Campinas-SP 13084-971, Brazil, wschwartz@liv.ic.unicam.br

Ricky J. Sethi University of California, Los Angeles, 4532 Boelter Hall, CA 90095-1596, USA, rickys@sethi.org; University of California, Los Angeles, 4532 Boelter Hall, CA 90095-1596, USA

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

Cristian Sminchisescu Institute for Numerical Simulation (INS), Faculty of Mathematics and Natural Science, University of Bonn, Bonn, Germany, cristian.sminchisescu@ins.uni-bonn.de; Institute for Mathematics of the Romanian Academy (IMAR), Bucharest, Romania

Bi Song University of California, Riverside, 900 University Ave. Riverside, CA 92521, USA, bsong@ee.ucr.edu

Ankur Srivastava Department of Electrical and Computer Engineering, and Institute for Systems Research, University of Maryland, College Park, MD 20742, USA, ankurs@umd.edu

Ming-Ting Sun University of Washington, Seattle, WA 98195, USA, mts@uw.edu

Graham Thomas BBC Research & Development, Centre House, 56 Wood Lane, London W12 7SB, UK, graham.thomas@bbc.co.uk

Fernando De la Torre Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA, ftorre@cs.cmu.edu

Cuong Tran Laboratory for Intelligent and Safe Automobiles (LISA), University of California at San Diego, San Diego, CA 92037, USA, cutran@ucsd.edu

Mohan Manubhai Trivedi Laboratory for Intelligent and Safe Automobiles (LISA), University of California at San Diego, San Diego, CA 92037, USA, mtrivedi@ucsd.edu

Alessandro Vinciarelli Computing Science Dept., University of Glasgow, Glasgow, UK, vincia@dcs.gla.acvi.uk; IDIAP Research Institute, Martigny, Switzerland

Xiaogang Wang Department of Electronic Engineering, Chinese University of Hong Kong, Hong Kong, China, xgwang@ee.cuhk.edu.hk

Tao Xiang Queen Mary University of London, London, E1 4NS, UK,
txiang@eecs.qmul.ac.uk

Zhanwu Xiong Computer Vision Center and Departament de Ciències de la Computació, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain,
zhanwu@cvc.uab.es

Part I
Detection and Tracking

Chapter 1
Is There Anybody Out There?

Thomas B. Moeslund

Applications within the field of Looking at People only make sense when the analyzed imagery contains one or more humans. The first step in such systems is therefore to determine if one or more humans are present and where in the scene they are. This task is termed *detection* or *figure-ground segmentation*. Moreover, since many applications require a number of consecutive frames containing people in order to do any processing, e.g., in many activity recognition tasks, *tracking* of individuals is often a requirement. This part provides an overview of detection and tracking methods. On one hand this part can be seen as the foundation on which the rest of the book builds, but detection and tracking methods are also applicable in their own rights, e.g., face detectors available in most new compact cameras, chroma-keying used for TV and movie productions, and in different surveillance applications.

In general, robust solutions to the detection and tracking problems are yet to be seen, but a massive effort can be observed by the number of papers published about these and related problems. Within the last decade or so a number of novel concepts and methods have been put forward allowing the research field to move further up the ladder toward the goal of systems that are able to work independently of, for example, illumination and weather conditions. Some of the more influential ones are the HoG detector [7], Viola and Jones' face detector [18] and the particle filter. These are discussed a bit more below and more thoroughly in the other chapters in Part I, and in general used in other systems discussed throughout this book.

Besides better methods, recent advancements can also be traced back to the introduction of public benchmarking data for assessing detection and tracking algorithms. This has first of all provided researchers and practitioners large annotated data sets to train and test their methods on, but equally important, those public data

T.B. Moeslund (✉)
Department of Architecture, Design and Media Technology, Aalborg University, Aalborg, Denmark
e-mail: tbm@create.aau.dk

T.B. Moeslund et al. (eds.), *Visual Analysis of Humans*,
DOI 10.1007/978-0-85729-997-0_1, © Springer-Verlag London Limited 2011

sets have allowed for different methods to be directly comparable, since they now can train and test on the same data sets. Some conferences even introduce competitions on data sets defined just for that event. Chapter 7 will give an overview of such data sets and how to evaluate them.

A last aspect that has to helped boost research and applications utilizing detection and tracking, is the fact that many of the methods have been implemented in software and made freely available. Good examples are OpenCV [6] aimed at engineers and computer scientists and EyesWeb [1] aimed at less mathematical and algorithmic oriented scholars. Both have equipped a whole generation of students and researchers from different fields with powerful tools for building Looking at People systems.

1.1 Detection

Two overall approaches to detecting people exist: pixel-based and object-based. In the former, each pixel in the incoming frame is compared to a model for that pixel in order to assess whether the incoming pixel is foreground or background. Having done so for the entire image a silhouette of each human in the incoming frame is (in theory) present. In the latter approach a sliding window is translated and scaled to all possible locations in the input frame and for each window the likelihood of it containing a human is calculated. This type of method will result in a bounding box (the window) containing the human as opposed to the silhouette resulting from the former method. These approaches are discussed further below—but first a few words on image acquisition.

1.1.1 Data Acquisition

Before any figure-ground segmentation can commence the frames need to be captured. Many algorithms require the humans in the frames to be of a reasonable resolution, otherwise the methods will fail. Combining this with the normal outdoor situation where a camera needs to cover a large scene renders the dilemma of resolution versus coverage. Increasing the field-of-view of the camera will provide better coverage, but lower resolution. Using cameras with a bigger image sensor can help, but this results in other problems like increased price. Another solution is to include multiple cameras and have them cooperate. This introduces the problem of handover during tracking, calibration among the cameras and of course increased price and data logistics. Yet another solution is to use an active sensor that can pan, tilt and zoom-in as need be. More of these might be cooperating and perhaps controlled in a master-and-slave fashion. No matter how this is organized, controlling active sensors is by no means trivial—especially when the object in focus (the human) can perform unpredicted movement. Chapter 2 is concerned with these issues and provides a more in-depth discussion. Moreover, some hands-on experiments are given in that chapter.

1.1.2 Pixel-Based Detection

The notion of having a model of the background and comparing each pixel in the incoming frame to that model is intuitively sound. The approach, however, has a major drawback assuming the background is fixed. While this works well in some indoor setting, it is in general not valid in outdoor scenes. Here trees will move in the wind and the illumination and shadows will change due to clouds and/or the shifting position of the sun. Current research therefore focuses on different ways of modeling the pixels in the background model and how to update such models during runtime. Especially the introduction of multiple models for each pixel [11, 13] has allowed for successful and real-time figure-ground segmentation in many applications. Pixel-based methods basically detect pixels from a new/moving object in the scene and hence some processing is required to determine whether the pixels are from a human, a car or something else. To this end filtering and blob analysis are normally required. Blob analysis can use shape cues to detect non-human-like objects, but the problem of shadows cast by humans is hard to solve since the shape of such blobs are naturally human-like. Different types of context-reasoning are therefore involved when trying to detect and delete shadows, for example the current weather conditions [8] or the fact that a shadow will be "bluish" since most of its illumination comes from the blue sky [12, 15]. Chapter 3 will provide more details on these matters.

1.1.3 Object-Based Detection

The pixel-based methods often fail in situations where the background is far from static, due to for example a moving camera, or when multiple people are occluding each other. To handle these situations the figure-ground segmentation problem can be addressed by using object-based detectors where the entire human (or major body parts) are detected directly. Such methods are often said to be window-based since they operate by translating a window over the input frame and calculating the likelihood of the window containing a human. Two methods have had a profound impact on such object-based approaches. The first is the HoG detector [7], which is built on the notion that different object shapes (here the human) always produce edges in an image and that these edges are not randomly distributed. Chapter 4 will describe how this, and other similar descriptors, can be used to detect humans even in complicated scenes.

Another significant approach to finding the human (or rather the face of a human) is the pioneering work behind Viola and Jones's face detector [18]. It combines simple features with the notion of cascaded classifiers. From studies into the human visual system it is known that some contrast detection is performed in specialized cells in the human eye. These rather simple operations have been simulated in computer vision using simple binary templates. Many different templates can be defined and a constellation of these can detect the face. The tricky issue is learning

which constellation of which templates that will do the job. Viola and Jones solved this complex learning problem by utilizing massive amount of positive and negative samples together with a cascade of simple classifiers. This idea was adapted from the field of machine learning and has afterwards been used in other computer vision subfields. In Chap. 5 Viola and Jones's face detector is introduced together with other issues related to face detection.

A major difference between the two approaches is that pixel-based methods detect whatever is moving, while object-based methods detect specific things using the prior knowledge about the foreground, hence the human. So, pixel-based methods require post processing, while object-based detectors can work on their own. Object-based detection methods produce a bounding box for each person in the frame. In contrast, pixel-based methods produce a silhouette for each person. What is preferred depends on the application. In controlled scenes, pixel-based methods work rather well, as seen in for example commercial chroma-keying systems, and can provide a very detailed segmentation and in a generally short processing time. But the object-based methods are in general better at detecting people in especially complicated scenes like outdoor settings with multiple occluding people and changing lighting conditions. The object-based methods are computational expensive, but with the introduction of for example GPU-based implementations this is less of a problem.

1.2 Tracking

Tracking is here defined as finding the temporal trajectory of an object through some state-space. The object would here often be the human but it could also be different body-parts as will be the case in Part II. The variables spanning the state-space are very often the 3D location parameters in the space or 2D locations on the image plane, but it could also be other parameters, e.g., color and shape. When tracking people we have in each a number of predictions and a number of new measurements. We need somehow to associate the measurements with the predictions in order to assign a tracking ID to each new measurement. This is in general known as the data association problem [5].

If we have a robust method for detecting people, then tracking is simply a matter of concatenating the output of the detector. This approach is known as tracking-by-detection and discussed in Chap. 4. Very often noisy and missing measurements will appear and the tracking-by-detection framework cannot stand alone. For the simple case where one person is tracked the Kalman filter framework has proven successful in combining predictions with noisy measurements. In the case of multiple people a multiple hypothesis approach will often form the tracking framework. These frameworks are seriously challenged in the case of unexpected events, such as new people entering the scene, people leaving the scene, people occluding each other, objects occluding people, and bad segmentation (false positive and false negative). To complicate matters even more, it is sometimes desired to track people

across non-overlapping cameras. A general solution to these problems is still far away.

Detecting people entering and leaving can to some extent be handled using the context of the scene, e.g., the fact that people do not just materialize or vanish, but tend to use doors. Knowing the trajectories of the past can help foresee the future, i.e. predicting where to search for people in the next frame. It might not be possible to track people during an occlusion, but the trajectories of the people after the occlusion can be compared with predicted trajectories to resolve ambiguities. More generally we can have short-term trajectory fragments with a low probability of error and view the tracking problem as a matter of merging these. Such trajectory fragments are denoted tracklets and described in more detail in Chap. 6 together with a method for tracking across multiple cameras.

If people stay occluded for some time even the notion of merging tracklets fails. Also, sometimes a system needs to track people during interaction. In these cases an appearance model of each individual is learned and updated online. Tracking can then be handled using a pixel-based segmentation method where each pixel in the input frame is compared with the different predicted appearance models of the different people. Alternatively, a generic tracker like template matching [9], mean-shift [16] or level sets [17] can be applied. This is discussed further in Chap. 4. When occlusion becomes a permanent situation, the individual is hard to track and instead we can analyze the movement patterns of the group or crowd the individual belongs to. This makes sense in applications where the objective is to understand for example the flow of people in airports or at public gatherings like concerts or sport events. Chapter 6 will discuss tracking in relation to groups and crowds.

1.3 Future Trends in Detection and Tracking

Pixel-based methods are point-based by nature and information about the neighbor pixels do not come into play before the post-processing stage. Random fields (MRF, CRF) or other approached to incorporate the spatial context is therefore an interesting approach to enhance foreground segmentation. Another possible use of context is to incorporate knowledge on the environment. If the 3D static environment and illumination sources could be modeled, then information about the spectral reflection properties of each surface could allow for a perfect computer graphics rendering of the current background. Combining this with dynamic information such as the current level and direction of sun's illumination would provide for a very robust pixel-based method. As mentioned above, pixel-based methods are fast and work well when the background can be modeled and updated. Moreover, during partial occlusion, pixels-based methods can also have their merits. On the other hand, object-based methods can operate without any scene knowledge, but are slower and tend to fail during occlusions. It seems only natural to combine these approaches as they can complement each other [17]. More of this can be expected in the future.

The sensor type plays a major role in detection and tracking. For example, a standard color camera will stand very little chance of detecting and tracking a human

in a pitch black scene, whereas a thermal camera will capture similar data no matter whether it is night or day. For exactly this reason infrared cameras (often with their own infrared lighting source) are becoming popular in surveillance scenarios and other applications where the detection can be solved in hardware or simple software. A good example of such an application is commercial motion capture equipment [4]. 3D sensing is another strategy that might play an important role in future acquisition systems. While different stereo solutions have been around for some time [14] a new type of compact 3D measurement devices are emerging, the time-of-flight cameras [2, 3]. They also provide 3D images of the scene, but using a more compact physical device. Data from 3D sensors can make detection a trivial task. The resolution and price of current time-of-flight cameras are still to be improved before becoming widely used in computer vision. Another 3D sensor with a much better resolution and lower price is the Kinect produced for Microsoft's Xbox. This technology is based on a structured light approach, where an infrared light pattern is cast onto the scene and picked up by a calibrated infrared camera. Such technology has a limited range of operation (usually less than 10 m) and requires that no other infrared light sources are present. But when these requirements are met the Kinect as such, and also the detecting and tracking software developed for the Xbox, seem like very good candidates for many looking at people systems. Current surveillance cameras often produce poor quality images due to issues like, low resolution, low frame-rate, poor colors and hard compression. Using better cameras and perhaps combining this with the new 3D capturing technologies is expected to help solve many of the ambiguity in both detection and tracking—especially in situations where occlusion is a problem.

Even though many data sets have been annotated and made publicly available, the detection and tracking communities still lack very long test data to see if the different algorithms can stand the test of time. It is very difficult to process a few minutes of videos and then conclude how the detector/tracker operates after being online 24/7/365. So far not many results on long sequences have been reported (1–2 days) [10] and results on longer periods of time are only evaluated qualitatively. What is needed is extremely long sequences, basically a whole year, to test the effects of the changing seasons.

References

1. http://www.infomus.org/eywmain.html [4]
2. http://www.mesa-imaging.ch/ [8]
3. http://www.pmdtec.com/ [8]
4. http://www.vicon.com [8]
5. Bar-Shalom, Y., Fortmann, T.E.: Tracking and Data Association. Academic Press, Boston (1988) [6]
6. Bradski, G., Kaehler, A.: Learning Opencv. O'Reilly Media Inc., Sebastopol (2008). http://oreilly.com/catalog/9780596516130 [4]
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Vision and Pattern Recognition (2005) [3,5]

8. Doshi, A., Trivedi, M.M.: Satellite imagery based robust, adaptive background models and shadow suppression. J. VLSJ Signal Process. **1**(2), 119–132 (2007) [5]
9. Ferrari, V., Marin, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (2008) [7]
10. Fihl, P., Corlin, R., Park, S., Moeslund, T.B., Trivedi, M.M.: Tracking of individuals in very long video sequences. In: International Symposium on Visual Computing, Lake Tahoe, Nevada, USA (2006) [8]
11. Grimson, W.E.L., Stauffer, C.: Adaptive background mixture models for real-time tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (1999) [5]
12. Huerta, I., Holte, M., Moeslund, T.B., Gonzàlez, J.: Detection and removal of chromatic moving shadows in surveillance scenarios. In: International Conference on Computer Vision, Kyoto, Japan (2009) [5]
13. Kim, K., Chalidabhongse, T.H., Harwood, D., Davis, L.: Background modeling and subtraction by codebook construction. In: International Conference on Image Processing (2004) [5]
14. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. **104**, 90–126 (2006) [8]
15. Prati, A., Mikic, I., Trivedi, M.M., Cucchiara, R.: Detecting moving shadows: Algorithms and evaluation. IEEE Trans. Pattern Anal. Mach. Intell. **25**(7), 918–923 (2003) [5]
16. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. In: ACM SIGGRAPH (2004) [7]
17. Stalder, S., Grabner, H., Van Gool, L.: Cascaded confidence filter for improved tracking-by-detection. In: European Conference on Computer Vision (2010) [7]
18. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Computer Vision and Pattern Recognition, pp. 511–518 (2001) [3,5]

Chapter 2
Beyond the Static Camera: Issues and Trends in Active Vision

Murad Al Haj, Carles Fernández, Zhanwu Xiong, Ivan Huerta, Jordi Gonzàlez, and Xavier Roca

Abstract Maximizing both the area coverage and the resolution per target is highly desirable in many applications of computer vision. However, with a limited number of cameras viewing a scene, the two objectives are contradictory. This chapter is dedicated to active vision systems, trying to achieve a trade-off between these two aims and examining the use of high-level reasoning in such scenarios. The chapter starts by introducing different approaches to active cameras configurations. Later, a single active camera system to track a moving object is developed, offering the reader first-hand understanding of the issues involved. Another section discusses practical considerations in building an active vision platform, taking as an example a multi-camera system developed for a European project. The last section of the chapter reflects upon the future trends of using semantic factors to drive smartly coordinated active systems.

M. Al Haj (✉) · C. Fernández · I. Huerta
Computer Vision Center, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain
e-mail: malhaj@cvc.uab.es

C. Fernández
e-mail: perno@cvc.uab.es

I. Huerta
e-mail: ivan.huerta@cvc.uab.es

Z. Xiong · J. Gonzàlez · X. Roca
Computer Vision Center and Departament de Ciències de la Computació, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain

Z. Xiong
e-mail: zhanwu@cvc.uab.es

J. Gonzàlez
e-mail: jordi.gonzalez@uab.cat

X. Roca
e-mail: xavier.roca@uab.cat

T.B. Moeslund et al. (eds.), *Visual Analysis of Humans*,
DOI 10.1007/978-0-85729-997-0_2, © Springer-Verlag London Limited 2011

2.1 Introduction

Many applications in the computer vision field benefit from high-resolution imagery. These include, but are not limited to, license-plate identification [4] and face recognition, where it has been observed that higher resolution improves accuracy [27]. For other applications, such as identifying people in surveillance videos, having highly zoomed images is a must. The problem with zoom control is that two opposing aims are desirable: the first one is obtaining a maximum resolution of the tracked object, whereas the second is minimizing the risk of losing this object. Therefore, zoom control can be thought of as a trade-off between the effective resolution per target and the desired coverage of the area of surveillance.

With a finite number of fixed sensors, there is a fundamental limit on the total area that can be observed. Thus, maximizing both the area of coverage and the resolution of each observed target requires an increase in the number of cameras. However, such an increase is highly costly in terms of installation and processing. Therefore, a system utilizing a smaller number of Pan–Tilt–Zoom (PTZ) cameras can be much more efficient if it is properly designed to overcome the obvious drawback of having less information about the target(s).

Toward this end, different works have investigated the use of PTZ cameras to address this problem of *actively* surveying a large area in an attempt to obtain high-quality imagery while maintaining coverage of the region [25]. Starting two decades ago, the area of active vision has been gaining much attention, in an attempt to: i) improve the quality of the acquired visual data by trying to keep a certain object at a desired scale, and ii) react to any changes in the scene dynamics that might risk the loss of the target.

Accurate reactive tracking of moving objects is a problem of both control and estimation. The speed at which the camera is adjusted must be a joint function of current camera position in pan, tilt and focal length, and the position of the tracked object in the 3D environment.

This chapter deals with active vision systems, offering the reader hands-on experience and insights into the problem. Section 2.2 discusses the different design alternatives for active cameras configurations, such as the autonomous camera approach, the master-slave approach and the active camera network approach, in addition to touching upon the advantages that environment reasoning lends to the problem. In Sect. 2.3, an autonomous camera system is designed, where the problem of jointly estimating the camera state and 3D object position is formulated as a Bayesian estimation problem and the joint state is estimated with an extended Kalman filter. The authors of this chapter had the opportunity to be part of a dedicated consortium working on a European project, called HERMES, where an integrated platform involving active cameras was built. Therefore, in Sect. 2.4, practical considerations involved in building real-time active camera systems are discussed taking the HERMES platform as a case study. This chapter is concluded in Sect. 2.5, where the lessons learned are summarized and the future directions are noted.

2.2 Active Camera Configurations

The interest in active camera systems started as early as two decades ago. Beginning in the late 1980s, Aloimonos et al. introduced the first general framework for active vision in order to improve the perceptual quality of tracking results [3]. Since then, numerous active camera systems have been developed. In this section, we take a look at different approaches for configuring these systems.

2.2.1 The Autonomous Camera Approach

Autonomous cameras are those that can self-direct in their surrounding environment. Recent work addressing this topic includes that of Denzler et al., where the motion of the tracked object is modeled using a Kalman filter. The camera focal length that minimizes the uncertainty in the state estimation is selected [12]. The authors used a stereo set-up, with two zoom cameras, to simplify the 3D estimation problem.

A newer approach is described by Tordoff et al., which tunes a constant velocity Kalman filter in order to ensure reactive zoom tracking while the focal length is varying [26]. Their approach correlates all the parameters of the filter with the focal length. However, they do not concentrate on the overall estimation problem, and their filter does not take into account any real-world object properties.

In the work by Nelson et al., a second rotating camera with fixed focal length is introduced in order to solve the problem of lost fixation [19].

The latter two works are primarily focused on zoom control and do not deal with total object-camera position estimation and its use in the control process. An attempt to join estimation and control in the same framework can be found in the work of Bagdanov et al., where a PTZ camera is used to actively track faces [5]. However, both the estimation and control models used are ad hoc, and the estimation approach is based on image features rather than 3D properties of the target being tracked.

2.2.2 The Master/Slave Approach

In a master/slave configuration, a supervising static camera is used to monitor a wide field of view and to track every moving target of interest. The position of each of these targets over time is then provided to a foveal camera, which tries to observe the targets at a higher resolution. Both the static and the active cameras are calibrated to a common reference, so that data coming from one of them can be easily projected onto the other, in order to coordinate the control of the active sensors.

Another possible use of the master/slave approach consists of a static (master) camera extracting visual features of an object of interest, while the active (slave) sensor uses these features to detect the desired object without the need of any training data. In this case, features should be invariant to illumination, viewpoint, color