Communications and Signal Processing – Vol. 4

Audio-Visual Person Tracking

A Practical Approach

Fotios Talantzis, Aristodemos Pnevmatikakis
and Anthony G Constantinides

Imperial College Press

Communications and Signal Processing – Vol. 4

Audio-Visual Person Tracking

A Practical Approach

Fotios Talantzis
Athens Information Technology, Greece
Imperial College London, UK

Aristodemos Pnevmatikakis
Athens Information Technology, Greece

Anthony G Constantinides
Imperial College London, UK

常州大学图书馆
藏书章

ICP Imperial College Press

Published by

Imperial College Press
57 Shelton Street
Covent Garden
London WC2H 9HE

Distributed by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

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

Communications and Signal Processing — Vol. 4
AUDIO VISUAL PERSON TRACKING
A Practical Approach

Copyright © 2012 by Imperial College Press

All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the Publisher.

For photocopying of material in this volume, please pay a copying fee through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. In this case permission to photocopy is not required from the publisher.

ISBN-13 978-1-84816-581-6
ISBN-10 1-84816-581-1

Printed in Singapore by Mainland Press Pte Ltd.

Audio-Visual Person Tracking

A Practical Approach

Communications and Signal Processing

Editors: Prof. A. Manikas & Prof. A. G. Constantinides
(Imperial College London, UK)

Vol. 1: Joint Source-Channel Coding of Discrete-Time Signals with Continuous Amplitudes
by Norbert Goertz

Vol. 2: Quasi-Orthogonal Space-Time Block Code
by Chau Yuen, Yong Liang Guan and Tjeng Thiang Tjhung

Vol. 3: Multi-Carrier Techniques for Broadband Wireless Communications: A Signal Processing Perspective
by C-C Jay Kuo, Michele Morelli and Man-On Pun

Vol. 4: Audio Visual Person Tracking: A Practical Approach
by Fotios Talantzis and Anthony G Constantinides

此为试读，需要完整PDF请访问：www.ertongbook.com

To my mother
F. Talantzis

To Efi, Athena and Katerina
A. Pnevmatikakis

To my students, past, present and future
A.G. Constantinides

Preface

Computing systems that are aware of human presence in order to provide heterogeneous services are gaining importance in living and working spaces, in entertainment, security and retail. A central role to such systems is the ability to sense humans and often track them in space across time. Tracking has become a mature topic in radar applications but requires a different set of sensors and algorithms when it involves humans. People generally do not like carrying tracking devices, a fact that inhibits service provision greatly. Instead, person tracking in this book is discussed in one of its unobtrusive flavours i.e. with the use of visual and audio modalities.

This book is about tracking humans using cameras and microphones, focusing on particle filtering algorithms. There are a few excellent texts on tracking [Blackman (1986); Blackman and Popoli (1999)], some of which focus on particle filters [Ristic *et al.* (2004)]. All these texts though focus on radar or sonar tracking. Audio-visual tracking needs different types of measurements on different types of signals: image [Gonzalez and Woods (2007)], video [Tekalp (1995); Forsyth and Ponce (2002); Shapiro and Stockman (2001)] and audio [Brandstein and Ward (2001)] signal processing elements need to be cast into the tracking frameworks. Two early works [Blake and Isard (1998); MacCormick (2002)] paved the way for visual tracking, but audio tracking still lacks a comprehensive text. A recent work covers audio-visual tracking [Zhu and Huang (2007)], mostly from the sensors and applications point of view.

Our aspiration is to fill in the gap between traditional tracking texts and signal processing texts. It is meant to be a solid introduction for the researcher starting in the field but also a good reference for people already working in it. It equips the reader with all the tools to measure the presence of humans in audio and visual signals and convert these measurements into likelihood functions. These likelihood functions are suitable for driving many types of tracking algorithms, but the emphasis is on particle filtering. This became an obvious choice after inspecting the evolution of the relevant literature in the past decade that slowly moved away from deterministic and Kalman versions to the more versatile framework of particle filters.

We believe that the coverage of the material is end-to-end, in the sense that the theoretical foundation of particle filtering and the necessary image, video, audio and array signal processing elements are first established, followed by working examples and MATLAB [Mathworks (2010); Gilat (2004)] implementations. The MATLAB implementations aim to serve as skeletons for the employment of novel systems. We felt that the book would not be complete without a chapter discussing applications and real-world systems. This allowed us to give a more meaningful aspect to an otherwise abstract scientific problem.

This book is written by two generations of authors: Two former students and their PhD supervisor. Thus, in numerous ways the completion of this book would not have been possible without the contribution of the supervisor in terms of guidance, inspiration and patience for over a decade. Additionally, the other two authors hope to have brought into the book a hands-on approach to a rather modern signal processing topic. Either way, throughout the composition, we all felt the obligation to create a book that will familiarise researchers with the topic and quickly enable them to further advance the algorithms.

This book project has not been an easy task. It is the consolidation of years of research in the field, some of it funded by European research projects such as Computers in the Human Interaction Loop [CHIL (2007)], Cognitive Care and Guidance for Active Ageing [HERMES (2010)] and Real-Time Context-Aware and Personalised Media Streaming Environments for Large Scale Broadcasting Applications [e Director 2012 (2010)]. It has been the return-on-investment after the supervision of numerous students, discussions with valuable colleagues and most importantly spending hours away from our families.

October 15, 2010
Fotios Talantzis,
Aristodemos Pnevmatikakis and
Antony G. Constantinides

Acknowledgments

The experience from many tracking systems is depicted in this book. All these systems have been implemented in the Autonomic and Grid Computing Lab of Athens Information Technology with the help of many people, either students (former or current) of the authors, or researchers collaborating with them. The authors wish to thank:

- Andreas Stergiou for the discussions held during endless implementations of visual trackers.
- Nikos Katsarakis for researching and implementing face detectors and particle filter visual and audio-visual trackers.
- Martin and Rasmus Andersen for their MSc theses resulting into a real-time multi-camera and multi-cue particle filter visual 3D tracker. Many thanks also to Prof. Zheng Hua Tan for co-supervising the two theses.
- Panos Papageorgiou for helping with the implementation of our very first visual particle filter tracker.
- Panos Kasianidis and Vasilis Mylonakis for their help in audio direction of arrival estimation systems.
- Ghassan Karame, Genadios Genaro and George Miggos for their implementations of 3D and audio-visual trackers during their MSc theses.
- Elias Rentzeperis, Thodoris Petsatodis and Christos Boukis for sharing their research in voice activity detection.
- Achilleas Anagnostopoulos for his thesis resulting in an augmented reality system with integrated palm tracking.
- Eric Lehmann for his valuable contribution in the early stages of the Audio Tracker.

- Nikos Dimakis and John Soldatos for the middleware design of the AIT SmartLab, that facilitates our live experiments and services.
- The partners in the CHIL, HERMES and My-e-Director 2012 projects (too numerous to list) for valuable discussions and opportunities for testing tracker systems in real application scenarios.
- Nikolaos Konstantinou for proof-reading parts of the manuscript and his suggestions.
- Finally, Nikos Vasilakis for designing the cover of this book.

Contents

Preface vii

Acknowledgments ix

List of Figures xv

List of Tables xxiii

1. Introduction 1

1.1 Person tracking 1
1.2 Why person tracking? 3
1.3 Person tracking challenges 4
1.4 Book organisation 5

2. Tracking Algorithms 9

2.1 Introduction 9
2.2 Tracking 10
2.3 Deterministic tracking 12
2.4 Stochastic trackers 13
2.4.1 Recursive Bayesian filtering 14
2.4.2 Kalman filter 16
2.4.3 Particle filter 23
2.5 Particle filter trackers 27
2.5.1 Object models 31
2.5.2 Object model for proposal distribution 33
2.5.3 Degeneracy of SIS particle filter 36
2.5.4 Optimum proposal distributions 37
2.5.5 Resampling 38

此为试读，需要完整PDF请访问：www.ertongbook.com

2.5.6 Sequential Importance Resampling particle filter . 40
2.6 Conclusions 43

3. Audio Tracking 45

3.1 Introduction 45
3.2 Multiple audio sensors 47
3.2.1 System model 48
3.2.2 The image model 49
3.2.3 Microphone arrays and their geometries 53
3.3 Audio trackers 56
3.3.1 Linear intersection 56
3.3.2 Particle filtering 58
3.3.3 Multiple speakers 61
3.4 Detection in audio 62
3.4.1 Time delay estimation 63
3.4.2 Beamforming 67
3.4.3 Voice activity detection 69
3.5 System parameters 72
3.5.1 Performance measures 72
3.5.2 Simulations 73

4. Visual Tracking 85

4.1 Introduction 85
4.2 From 3D world to 2D images 86
4.3 3D imaging 91
4.4 Measurements in images 95
4.5 Colour matching 96
4.5.1 Colour without spatial information 97
4.5.2 Colour with spatial information 101
4.6 Outline matching 106
4.6.1 Clutter attenuation by edge orientation 111
4.6.2 View changes handling by modelling in 3D 113
4.7 Moving objects 119
4.7.1 Frame-by-frame difference 119
4.7.2 Adaptive background estimation 120
4.7.3 Motion likelihood 121
4.8 Face detection 122
4.8.1 Viola–Jones face detector 123

4.8.2 Face validation 124
4.8.3 Face likelihood 128
4.9 Visual tracking systems 130
4.10 Deterministic tracking 132
4.10.1 Template matching 133
4.10.2 Mean-Shift 133
4.11 Stochastic tracking 135
4.11.1 Kalman filter 135
4.11.2 Colour-based particle filter face tracker 139
4.11.3 Measurement-assisted proposal distribution . . . 146
4.11.4 Particle filter face tracker with colour-assisted proposal distribution 147
4.11.5 Particle filter face tracker using face likelihood . . 150
4.12 Conclusions 154

5. Multimodal Tracking 155

5.1 Introduction 155
5.2 Likelihood combination 156
5.2.1 Multiplying face and colour likelihoods 157
5.3 Tracker output combination 160
5.3.1 Deterministic fusion of independent trackers . . . 160
5.3.2 Kalman fusion of independent audio and visual trackers 164
5.4 Partitioned sampling 166
5.4.1 Joint multi-cue face tracker employing partitioned sampling 167
5.4.2 Joint audio and visual tracker employing partitioned sampling 170
5.5 Sensor synchronisation 172
5.6 Conclusions 173

6. Applications 175

6.1 Smart Spaces 175
6.1.1 The Smart Home 176
6.1.2 The Smart Office 181
6.1.3 Video-conferencing 185
6.2 Entertainment 187
6.3 Personalised media broadcasting 188

6.4 Security and surveillance 190
6.5 Robotics 192
6.6 Marketing and retail 194
6.7 Conclusions 196

Bibliography 197

Index 207

List of Figures

2.1 Deterministic tracking as a two-step process. The state at the previous time instance is propagated according to the object model to the one-step prediction. A grid search is conducted around this prediction in the state-space, by mapping all grid points to regions in the signal space (here image plane) and measuring for best match using the measurement model. The best match is mapped back to the state-space, yielding the current state estimation. 13

2.2 1D example of Kalman filtering. One-step prediction as deterministic drift of the conditional state mean and uncertainty increase, and measurement update by moving the prediction towards the measurement and decreasing the uncertainty based on the Kalman gain. Note that to have the conditional means and the measurement on the same axis, the former needs to be multiplied by the measurement model matrix, which is omitted for simplicity. 17

2.3 2D tracks from example 2.2, for three values of the ratio of the measurement over the process noise variances. Only the position is shown for the hidden states (actual locations), the state estimates and the one-step predictions. For the latter, dotted lines begin from the previous state estimate and finish at the predictions. A second dotted line connects the predictions with the current estimate based on the measurement update. . 22

2.4 RMS error between the estimated and the hidden states as a function of the ratio of the measurement over the process noise variances for example 2.2. 23

2.5 Representation of a distribution with particles. The particles are depicted as disks, with centres corresponding to the sample values and radii proportional to the weights of the particles. . . 24

2.6 Importance sampling of a bimodal target distribution using a wide Gaussian proposal. 26

2.7 Target and proposal distributions for the importance sampling of example 2.4. 27

2.8 RMSE of the estimation of the mean value of the target distribution for the three different proposal distributions of example 2.4 (see Fig. 2.7) as a function of the number N_p of particles used. 28

2.9 One iteration of the SIS particle filter, involving drawing new samples from the proposal distribution conditioned upon the previous samples and the current measurement, and the update of their weights based on (2.48). 31

2.10 Bimodal PDF of the form of the object model in (2.49), estimated from the samples drawn as discussed in example 2.5. The uniform and the Gaussian components are clearly visible. . 33

2.11 Particles of the SIS particle filter in example 2.6 for the first few time steps. The hidden states to be estimated are shown in red circles, the observable measurements in blue squares and the estimated states in green "x". The filter estimates the hidden states better than the measurements. 36

2.12 Systematic resampling of the original sample values $\mathbf{x}^{(j)}$ into $\mathbf{x}_r^{(i)}$, shown at the horizontal axes. The vertical axis depicts the cumulative sum of the weights, which is resampled at regular intervals $1/N_p$, starting at the random choice u_1. 39

2.13 RMSE of the standard deviation estimation in example 2.7, with and without resampling, as a function of the number of particles used. 41

2.14 Operation of the SIR particle filter, including drawing new samples from the object model, updating the weights according to (2.50) and resampling to avoid the degeneracy problem. 42

2.15 Median and inter-quartile range of the RMS state estimation error of the particle filter of example 2.6 for 40 time steps, without (dashed lines) and with resampling (solid lines). The latter is the SIR particle filter. The inter-quartile range is represented by the length of the vertical bars centred on the median. 42

2.16 Particles of the SIR particle filter in example 2.6 for the first few time steps. The hidden states to be estimated are shown in red circles, the observable measurements in blue squares and the estimated states in green "x". To be compared to the trajectories without resampling in Fig. 2.16. 43

3.1 Block diagram of an ASLT system. 47
3.2 Schematic representation of the image model. 50
3.3 Array topologies schematically. 54
3.4 Geometry of a single pair. 56
3.5 Quadruple sensor arrangement and local Cartesian coordinate system. 57
3.6 Flow diagram of hangover scheme. 71
3.7 Overview of simulated room. 75
3.8 Effect of reverberation time T_{60}. Average RMSE over 30 simulations is shown for all algorithms. 76
3.9 Effect of reverberation time T_{60}. Average MSD over 30 simulations is shown for all algorithms. 77
3.10 Effect of order $\mathcal{N}$. Average RMSE over 30 simulations. 79
3.11 Effect of the VAD module. Average RMSE over 30 simulations. 80
3.12 Effect of the array topology. Average RMSE over 30 simulations. 81
3.13 Effect of interchanging sources. TVRMSE over 30 simulations is shown for the MI-TDE and GCC algorithms. 82
3.14 Effect of interchanging sources. TVRMSE over 30 simulations is shown for the MI-BF and BF algorithms. 83

4.1 Viewing a scene with a camera to project it onto an image plane. Three coordinate systems are involved. 87
4.2 Projection of the head centres onto the five camera views for example 4.1. 90
4.3 Viewing a scene with a camera to project it onto an image plane. Three coordinate systems are involved. 92
4.4 Data-driven 3D tracking by combination of the independent 2D face tracks (red rectangles) from multiple views that correspond to the same target. After combination, the 3D coordinates obtained using `locate3D()` are projected onto the floor (height is ignored) and are shown as red "x" marks. 94
4.5 Image and associated colour likelihood with a generic human skin colour histogram as reference. 100

此为试读，需要完整PDF请访问：www.ertongbook.com