Shaogang Gong
Tao Xiang

Visual Analysis of Behaviour

From Pixels to Semantics

Springer

Visual Analysis of Behaviour

Shaogang Gong • Tao Xiang

Visual Analysis of Behaviour

From Pixels to Semantics

Springer

Shaogang Gong
School of Electronic Engineering
and Computer Science
Queen Mary University of London
Mile End Rd.
London, E1 4NS
UK
sgg@eecs.qmul.ac.uk

Tao Xiang
School of Electronic Engineering
and Computer Science
Queen Mary University of London
Mile End Rd.
London, E1 4NS
UK
txiang@eecs.qmul.ac.uk

ISBN 978-0-85729-669-6 e-ISBN 978-0-85729-670-2
DOI 10.1007/978-0-85729-670-2
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: 2011929785

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)

To Aleka, Philip and Alexander
Shaogang Gong

To Ning and Rachel
Tao Xiang

Preface

The human visual system is able to visually recognise and interpret object behaviours under different conditions. Yet, the goal of building computer vision based recognition systems with comparable capabilities has proven to be very difficult to achieve. Computational modelling and analysis of object behaviours through visual observation is inherently ill-posed. Many would argue that our cognitive understanding remains unclear about why we associate certain semantic meanings with specific object behaviours and activities. This is because meaningful interpretation of a behaviour is subject to the observer's a priori knowledge, which is at times rather ambiguous. The same behaviour may have different semantic meanings depending upon the context within which it is observed. This ambiguity is exacerbated when many objects are present in a scene. Can a computer based model be constructed that is able to extract all necessary information for describing a behaviour from visual observation alone? Do people behave differently in the presence of others, and if so, how can a model be built to differentiate the expected normal behaviours from those of abnormality? Actions and activities associated with the same behavioural interpretation may be performed differently according to the intended meaning, and different behaviours may be acted in a subtly similar way. The question arises as to whether these differences can be accurately measured visually and robustly computed consistently for meaningful interpretation of behaviour.

Visual analysis of behaviour requires not only to solve the problems of object detection, segmentation, tracking, motion trajectory analysis, but also the modelling of context information and utilisation of non-sensory knowledge when available, such as human annotation of input data or relevance feedback to output signals. Visual analysis of behaviour faces two fundamental challenges in computational complexity and uncertainty. Object behaviours in general exhibit complex spatio-temporal dynamics in a highly dynamical and uncertain environment, for instance, human activities in a crowded public space. Segmenting and modelling human actions and activities in a visual environment is inherently ill-posed, as information processing in visual analysis of behaviour is subject to noise, incompleteness and uncertainty in sensory data. Whilst these visual phenomena are difficult to model analytically, they can be probabilistically modelled much more effectively through statistical machine learning.

Despite these difficulties, it is compelling that one of the most significant developments in computer vision research over the last 20 years has been the rapidly growing interest in automatic visual analysis of behaviour in video data captured from closed-circuit television (CCTV) systems installed in private and public spaces. The study of visual analysis of behaviour has had an almost unique impact on computer vision and machine learning research at large. It raises many challenges and provides a testing platform for examining some difficult problems in computational modelling and algorithm design. Many of the issues raised are relevant to dynamic scene understanding in general, multivariate time series analysis and statistical learning in particular.

Much progress has been made since the early 1990s. Most noticeably, statistical machine learning has become central to computer vision in general, and to visual analysis of behaviour in particular. This is strongly reflected throughout this book as one of the underlying themes. In this book, we study plausible computational models and tractable algorithms that are capable of automatic visual analysis of behaviour in complex and uncertain visual environments, ranging from well-controlled private spaces to highly crowded public scenes. The book aims to reflect the current trends, progress and challenges on visual analysis of behaviour. We hope this book will not only serve as a sampling of recent progress but also highlight some of the challenges and open questions in automatic visual analysis of object behaviour.

There is a growing demand by both governments and commerce worldwide for advanced imaging and computer vision technologies capable of automatically selecting and identifying behaviours of objects in imagery data captured in both public and private spaces for crime prevention and detection, public transport management, personalised healthcare, information management and market studies, asset and facility management. A key question we ask throughout this book is how to design automatic visual learning systems and devices capable of extracting and mining salient information from vast quantity of data. The algorithm design characteristics of such systems aim to provide, with minimum human intervention, machine capabilities for extracting relevant and meaningful semantic descriptions of salient objects and their behaviours for aiding decision-making and situation assessment.

There have been several books on human modelling and visual surveillance over the years, including *Face Detection and Gesture Recognition for Human-Computer Interaction* by Yang and Ahuja (2001); *Analyzing Video Sequences of Multiple Humans* by Ohya, Utsumi and Yamato (2002); *A Unified Framework for Video Summarization, Browsing and Retrieval: with Applications to Consumer and Surveillance Video* by Xiong, Radhakrishnan, Divakaran, Rui and Huang (2005); *Human Identification based on Gait* by Nixon, Tan and Chellapa (2005); and *Automated Multi-Camera Surveillance* by Javed and Shah (2008). There are also a number of books and edited collections on behaviour studies from cognitive, social and psychological perspectives, including *Analysis of Visual Behaviour* edited by Ingle, Goodale and Mansfield (1982); *Hand and Mind: What Gestures Reveal about Thought* by McNeill (1992); *Measuring Behaviour* by Martin (1993); *Understanding Human Behaviour* by Mynatt and Doherty (2001); and *Understanding Human Behaviour and the Social Environment* by Zastrow and Kirst-Ashman (2003). However, there has been no book that provides a comprehensive and unified treatment of visual analysis of behaviour from a computational modelling and algorithm design perspective.

This book has been written with an emphasis on computationally viable approaches that can be readily adopted for the design and development of intelligent computer vision systems for automatic visual analysis of behaviour. We present what is fundamentally a computational algorithmic approach, founded on recent advances in visual representation and statistical machine learning theories. This approach should also be attractive to researchers and system developers who would like to both learn established techniques for visual analysis of object behaviour, and gain insight into up-to-date research focus and directions for the coming years. We hope that this book succeeds in providing such a treatment of the subject useful not only for the academic research communities, both also the commerce and industry.

Overall, the book addresses a broad range of behaviour modelling problems, from established areas of human facial expression, body gesture and action analysis to emerging new research topics in learning group activity models, unsupervised behaviour profiling, hierarchical behaviour discovery, learning behavioural context, modelling rare behaviours, 'man-in-the-loop' active learning of behaviours, multi-camera behaviour correlation, person re-identification, and 'connecting-the-dots' for global abnormal behaviour detection. The book also gives in depth treatment to some popular computer vision and statistical machine learning techniques, including the Bayesian information criterion, Bayesian networks, 'bag-of-words' representation, canonical correlation analysis, dynamic Bayesian networks, Gaussian mixtures, Gibbs sampling, hidden conditional random fields, hidden Markov models, human silhouette shapes, latent Dirichlet allocation, local binary patterns, locality preserving projection, Markov processes, probabilistic graphical models, probabilistic topic models, space-time interest points, spectral clustering, and support vector machines.

The computational framework presented in this book can also be applied to modelling behaviours exhibited by many other types of spatio-temporal dynamical systems, either in isolation or in interaction, and therefore can be beneficial to a wider range of fields of studies, including Internet network behaviour analysis and profiling, banking behaviour profiling, financial market analysis and forecasting, bioinformatics, and human cognitive behaviour studies.

We anticipate that this book will be of special interest to researchers and academics interested in computer vision, video analysis and machine learning. It should be of interest to industrial research scientists and commercial developers keen to exploit this emerging technology for commercial applications including visual surveillance for security and safety, information and asset management, public transport and traffic management, personalised healthcare in assisting elderly and disabled, video indexing and search, human computer interaction, robotics, animation and computer games. This book should also be of use to post-graduate students of computer science, mathematics, engineering, physics, behavioural science, and cognitive psychology. Finally, it may provide government policy makers and commercial managers an informed guide on the potentials and limitations in deploying intelligent video analytics systems.

The topics in this book cover a wide range of computational modelling and algorithm design issues. Some knowledge of mathematics would be useful for the reader. In particular, it would be convenient if one were familiar with vectors and matrices, eigenvectors and eigenvalues, linear algebra, optimisation, multivariate analysis, probability, statistics and calculus at the level of post-graduate mathematics. However, the non-mathematically inclined reader should be able to skip over many of the equations and still understand much of the content.

London, UK

Shaogang Gong
Tao Xiang

Acknowledgements

We shall express our deep gratitude to the many people who have helped us in the process of writing this book. The experiments described herein would not have been possible without the work of PhD students and postdoctoral research assistants at Queen Mary University of London. In particular, we want to thank Chen Change Loy, Tim Hospedales, Jian Li, Wei-Shi Zheng, Jianguo Zhang, Caifeng Shan, Yogesh Raja, Bryan Prosser, Matteo Bregonzio, Parthipan Siva, Lukasz Zalewski, Eng-Jon Ong, Jamie Sherrah, Jeffrey Ng, and Michael Walter for their contributions to this work. We are indebted to Alexandra Psarrou who read a draft carefully and gave us many helpful comments and suggestions.

We shall thank Simon Rees and Wayne Wheeler at Springer for their kind help and patience during the preparation of this book. The book was typeset using LaTeX.

We gratefully acknowledge the financial support that we have received over the years from UK EPSRC, UK DSTL, UK MOD, UK Home Office, UK TSB, US Army Labs, EU FP7, the Royal Society, BAA, and QinetiQ. Finally, we shall thank our families and friends for all their support.

Contents

Part I Introduction

1 About Behaviour 3
1.1 Understanding Behaviour 4
1.1.1 Representation and Modelling 5
1.1.2 Detection and Classification 6
1.1.3 Prediction and Association 6
1.2 Opportunities 7
1.2.1 Visual Surveillance 8
1.2.2 Video Indexing and Search 8
1.2.3 Robotics and Healthcare 9
1.2.4 Interaction, Animation and Computer Games 9
1.3 Challenges 10
1.3.1 Complexity 10
1.3.2 Uncertainty 10
1.4 The Approach 11
References 12

2 Behaviour in Context 15
2.1 Facial Expression 15
2.2 Body Gesture 16
2.3 Human Action 19
2.4 Human Intent 21
2.5 Group Activity 22
2.6 Crowd Behaviour 24
2.7 Distributed Behaviour 25
2.8 Holistic Awareness: Connecting the Dots 28
References 29

3 Towards Modelling Behaviour 39
3.1 Behaviour Representation 39

3.1.1 Object-Based Representation 39
3.1.2 Part-Based Representation 43
3.1.3 Pixel-Based Representation 44
3.1.4 Event-Based Representation 46
3.2 Probabilistic Graphical Models 48
3.2.1 Static Bayesian Networks 50
3.2.2 Dynamic Bayesian Networks 50
3.2.3 Probabilistic Topic Models 52
3.3 Learning Strategies 53
3.3.1 Supervised Learning 53
3.3.2 Unsupervised Learning 54
3.3.3 Semi-supervised Learning 55
3.3.4 Weakly Supervised Learning 56
3.3.5 Active Learning 56
References 58

Part II Single-Object Behaviour

4 Understanding Facial Expression 69
4.1 Classification of Images 69
4.1.1 Local Binary Patterns 70
4.1.2 Designing Classifiers 73
4.1.3 Feature Selection by Boosting 76
4.2 Manifold and Temporal Modelling 78
4.2.1 Locality Preserving Projections 78
4.2.2 Bayesian Temporal Models 85
4.3 Discussion 88
References 90

5 Modelling Gesture 95
5.1 Tracking Gesture 95
5.1.1 Motion Moment Trajectory 95
5.1.2 2D Colour-Based Tracking 97
5.1.3 Bayesian Association 99
5.1.4 3D Model-Based Tracking 109
5.2 Segmentation and Atomic Action 115
5.2.1 Temporal Segmentation 115
5.2.2 Atomic Actions 117
5.3 Markov Processes 119
5.4 Affective State Analysis 123
5.4.1 Space-Time Interest Points 124
5.4.2 Expression and Gesture Correlation 126
5.5 Discussion 128
References 128

6 Action Recognition 133
6.1 Human Silhouette 134
6.2 Hidden Conditional Random Fields 135
6.2.1 HCRF Potential Function 138
6.2.2 Observable HCRF 138
6.3 Space-Time Clouds 141
6.3.1 Clouds of Space-Time Interest Points 141
6.3.2 Joint Local and Global Feature Representation 149
6.4 Localisation and Detection 150
6.4.1 Tracking Salient Points 151
6.4.2 Automated Annotation 154
6.5 Discussion 157
References 158

Part III Group Behaviour

7 Supervised Learning of Group Activity 163
7.1 Contextual Events 164
7.1.1 Seeding Event: Measuring Pixel-Change-History 164
7.1.2 Classification of Contextual Events 167
7.2 Activity Segmentation 170
7.2.1 Semantic Content Extraction 170
7.2.2 Semantic Video Segmentation 175
7.3 Dynamic Bayesian Networks 181
7.3.1 Correlations of Temporal Processes 181
7.3.2 Behavioural Interpretation of Activities 185
7.4 Discussion 190
References 191

8 Unsupervised Behaviour Profiling 193
8.1 Off-line Behaviour Profile Discovery 194
8.1.1 Behaviour Pattern 194
8.1.2 Behaviour Profiling by Data Mining 195
8.1.3 Behaviour Affinity Matrix 196
8.1.4 Eigendecomposition 196
8.1.5 Model Order Selection 197
8.1.6 Quantifying Eigenvector Relevance 198
8.2 On-line Anomaly Detection 200
8.2.1 A Composite Behaviour Model 200
8.2.2 Run-Time Anomaly Measure 203
8.2.3 On-line Likelihood Ratio Test 204
8.3 On-line Incremental Behaviour Modelling 205
8.3.1 Model Bootstrapping 206
8.3.2 Incremental Parameter Update 208
8.3.3 Model Structure Adaptation 210
8.4 Discussion 211
References 212

9 Hierarchical Behaviour Discovery . . . 215
9.1 Local Motion Events . . . 216
9.2 Markov Clustering Topic Model . . . 217
9.2.1 Off-line Model Learning by Gibbs Sampling . . . 219
9.2.2 On-line Video Saliency Inference . . . 221
9.3 On-line Video Screening . . . 223
9.4 Model Complexity Control . . . 226
9.5 Semi-supervised Learning of Behavioural Saliency . . . 228
9.6 Discussion . . . 230
References . . . 231

10 Learning Behavioural Context . . . 233
10.1 Spatial Context . . . 235
10.1.1 Behaviour-Footprint . . . 235
10.1.2 Semantic Scene Decomposition . . . 236
10.2 Correlational and Temporal Context . . . 238
10.2.1 Learning Regional Context . . . 239
10.2.2 Learning Global Context . . . 243
10.3 Context-Aware Anomaly Detection . . . 245
10.4 Discussion . . . 247
References . . . 248

11 Modelling Rare and Subtle Behaviours . . . 251
11.1 Weakly Supervised Joint Topic Model . . . 252
11.1.1 Model Structure . . . 252
11.1.2 Model Parameters . . . 255
11.2 On-line Behaviour Classification . . . 262
11.3 Localisation of Rare Behaviour . . . 263
11.4 Discussion . . . 264
References . . . 265

12 Man in the Loop . . . 267
12.1 Active Behaviour Learning Strategy . . . 269
12.2 Local Block-Based Behaviour . . . 271
12.3 Bayesian Classification . . . 273
12.4 Query Criteria . . . 274
12.4.1 Likelihood Criterion . . . 275
12.4.2 Uncertainty Criterion . . . 275
12.5 Adaptive Query Selection . . . 277
12.6 Discussion . . . 279
References . . . 282

Part IV Distributed Behaviour

13 Multi-camera Behaviour Correlation . . . 285
13.1 Multi-view Activity Representation . . . 288
13.1.1 Local Bivariate Time-Series Events . . . 288

13.1.2 Activity-Based Scene Decomposition 289
13.2 Learning Pair-Wise Correlation 292
13.2.1 Cross Canonical Correlation Analysis 293
13.2.2 Time-Delayed Mutual Information Analysis 295
13.3 Multi-camera Topology Inference 296
13.4 Discussion 298
References 298

14 Person Re-identification 301
14.1 Re-identification by Ranking 303
14.1.1 Support Vector Ranking 303
14.1.2 Scalability and Complexity 305
14.1.3 Ensemble RankSVM 306
14.2 Context-Aware Search 308
14.3 Discussion 310
References 312

15 Connecting the Dots 315
15.1 Global Behaviour Segmentation 315
15.2 Bayesian Behaviour Graphs 318
15.2.1 A Time-Delayed Probabilistic Graphical Model 319
15.2.2 Bayesian Graph Structure Learning 321
15.2.3 Bayesian Graph Parameter Learning 326
15.2.4 Cumulative Anomaly Score 326
15.2.5 Incremental Model Structure Learning 329
15.3 Global Awareness 334
15.3.1 Time-Ordered Latent Dirichlet Allocation 334
15.3.2 On-line Prediction and Anomaly Detection 337
15.4 Discussion 339
References 340

Epilogue 343

Index 345

Acronyms

1D	one-dimensional
2D	two-dimensional
3D	three-dimensional
BIC	Bayesian information criterion
CCA	canonical correlation analysis
CCTV	closed-circuit television
CONDENSATION	conditional density propagation
CRF	conditional random field
EM	expectation-maximisation
DBN	dynamic Bayesian network
HCI	human computer interaction
HCRF	hidden conditional random field
HOG	histogram of oriented gradients
FACS	facial action coding system
FOV	field of view
FPS	frame per second
HMM	hidden Markov model
KL	Kullback–Leibler
LBP	local binary patterns
LDA	latent Dirichlet allocation
LPP	locality preserving projection
MAP	maximum a posteriori
MCMC	Markov chain Monte Carlo
MLE	maximum likelihood estimation
MRF	Markov random field
PCA	principal component analysis
PGM	probabilistic graphical model
PTM	probabilistic topic model
PTZ	pan-tilt-zoom
SIFT	scale-invariant feature transform
SLPP	supervised locality preserving projection
SVM	support vector machine
xCCA	cross canonical correlation analysis

Part I
Introduction

Chapter 1
About Behaviour

Understanding and interpreting behaviours of objects, and in particular those of humans, is central to social interaction and communication. Commonly, one considers that behaviours are the actions and reactions of a person or animal in response to external or internal stimuli. There is, however, a plethora of wider considerations of what behaviour is, ranging from economical (Simon 1955), organisational (Rollinson 2004), social (Sherman and Sherman 1930), to sensory attentional such as *visual behaviour* (Ingle et al. 1982). Visual behaviour refers to the actions or reactions of a sensory mechanism in response to a visual stimulus, for example, the navigation mechanism of nocturnal bees in dim light (Warrant 2008), visual search by eye movement of infants (Gough 1962) or drivers in response to their surrounding environment (Harbluk and Noy 2002). If visual behaviour as a search mechanism is a perceptual function that scans actively a visual environment in order to focus attention and seek an object of interest among distracters (Ltti et al. 1998), *visual analysis of behaviour* is a perceptual task that interprets actions and reactions of objects, such as people, interacting or co-existing with other objects in a visual environment (Buxton and Gong 1995; Gong et al. 2002; Xiang and Gong 2006). The study of visual analysis of behaviour, and in particular of human behaviour, is the focus of this book.

Recognising objects visually by behaviour and activity rather than shape and size plays an important role in a primate visual system (Barbur et al. 1980; Schiller and Koerner 1971; Weiskrantz 1972). In a visual environment of multiple objects co-existing and interacting, it becomes necessary to identify objects not only by their appearance but also by what they do. The latter provides richer information about objects especially when visual data are spatially ambiguous and incomplete. For instance, some animals, such as most snakes, have a very poor visual sensing system, which is unable to capture sufficient visual appearance of objects but very sensitive to movements for detecting preys and predators. The human visual system is highly efficient for scanning through large quantity of low-level imagery data and selecting salient information for a high-level semantic interpretation and gaining situational awareness.

S. Gong, T. Xiang, *Visual Analysis of Behaviour*,
DOI 10.1007/978-0-85729-670-2_1, © Springer-Verlag London Limited 2011

1.1 Understanding Behaviour

Since 1970s, the computer vision community has endeavoured to bring about intelligent perceptual capabilities to artificial visual sensors. Computer vision aims to build artificial mechanisms and devices capable of mimicking the sensing capabilities of biological vision systems (Marr 1982). This endeavour is intensified in recent years by the need for understanding massive quantity of video data, with the aim to not only comprehend objects spatially in a snapshot but also their spatio-temporal relations over time in a sequence of images. For understanding a dynamically changing social environment, a computer vision system can be designed to interpret behaviours from object actions and interactions captured visually in that environment. A significant driver for visual analysis of behaviour is automated visual surveillance, which aims to automatically interpret human activities and detect unusual events that could pose a threat to public security and safety.

If a behaviour is considered the way how an object acts, often in relation to other objects in the same visual environment, the focus of this book is on visual analysis of human behaviour and behaviours of object that are manipulated by humans, for example, vehicles driven by people. There are many interchangeable terms used in the literature concerning behaviour, including activities, actions, events, and movements. They correspond to different spatial and temporal context within which a behaviour can be defined. One may consider a behaviour hierarchy of three layers:

1. Atomic actions correspond to instantaneous atomic entities upon which an action is formed. For example, in a running action, the atomic action could be 'left leg moving in front of the right leg'. In a returning action in tennis, it could be 'swing right hand' followed by 'rotating the upper body'.
2. Actions correspond to a sequence of atomic actions that fulfil a function or purpose. For instance, walking, running, or serving a tennis ball.
3. Activities are composed of sequences of actions over space and time. For example, 'a person walking from a living room to a kitchen to fetch a cup of water', or 'two people playing tennis'. Whilst actions are likely associated with a single object in isolation, activities are almost inevitably concerned with either interactions between objects, or an object engaging with the surrounding environment.

In general, visual analysis of behaviour is about constructing models and developing devices for automatic analysis and interpretation of object actions and activities captured in a visual environment. To that end, visual analysis of behaviour focuses on three essential functions:

1. Representation and modelling: To extract and encode visual information from imagery data in a more concise form that also captures intrinsic characteristics of objects of interest;
2. Detection and classification: To discover and search for salient, perhaps also unique, characteristics of certain object behaviour patterns from large quantity of visual observations, and to discriminate them against known categories of semantic and meaningful interpretation;

3. Prediction and association: To forecast future events based on the past and current interpretation of behaviour patterns, and to forge object identification through behavioural expectation and trend.

We consider that automated visual analysis of behaviour is information processing of visual data, capable of not only modelling previously observed object behaviours, but also detecting, recognising and predicting unseen behavioural patterns and associations.

1.1.1 Representation and Modelling

A human observer can recognise behaviours of interest directly from visual observation. This suggests that imagery data embed useful information for semantic interpretation of object behaviour. Behaviour representation addresses the question of what information must be extracted from images and in what form, so that object behaviour can be understood and recognised visually. The human visual system utilises various visual cues and contextual information for recognising objects and their behaviour (Humphreys and Bruce 1989). For instance, the specific stripe pattern and its colour is a useful cue for human to recognise a tiger and distinguish it from other cats, such as lions. Similarly, the movement as well as posture of a tiger can reveal its intended action: running, walking, or about to strike. It is clear that different sources and types of visual information need be utilised for modelling and understanding object behaviour.

A behaviour representation needs to accommodate both cumulative and temporal information about an object. In order to recognise an object and its behaviour, the human visual system relates any visual stimuli falling on to the retina to a set of knowledge and expectation about the object under observation: how it *should* look like and how it is *supposed* to behave (Gregory 1970; von Helmholtz 1962). Behaviour representation should address the need for extracting visual information that can facilitate the association of visual observation with semantic interpretation. In other words, representation of visual data is part of a computational mechanism that contributes towards constructing and accumulating knowledge about behaviour. For example, modelling the action of a person walking can be considered as to learn the prototypical and generic knowledge of walking based on limited observations of walking examples, so that when an unseen instance of walking is observed, it can be recognised by utilising the accumulated a priori knowledge. An important difference between object modelling and behaviour modelling is that a behaviour model should benefit more from capturing temporal information about behaviour. Object recognition in large only considers spatial information. For instance, a behaviour model is built based on visual observation of a person's daily routine in an office which consists of meetings, tea breaks, paper works and a lunch at certain times of every day. What has been done so far can then have a significant influence on the correct interpretation of what this person is about to do (Agre 1989).

A computational model of behaviour performs both representation and matching. For representing object behaviour, one considers a model capable of capturing distinctive characteristics of an object in action and activity. A good behaviour representation aims to describe an object sufficiently well for both generalisation and discrimination in model matching. Model matching is a computational process to either explain away new instances of observation against known object behaviours, considered as its generalisation capacity, or discriminate one type of object behaviour from the others, regarded as its discrimination ability. For effective model matching, a representation needs to separate visual observation of different object behaviour types or classes in a representational space, and to maintain such separations given noisy and incomplete visual observations.

1.1.2 Detection and Classification

Generally speaking, visual classification is a process of categorising selected visual observations of interest into known classes. Classification is based on an assumption that segmentation and selection of interesting observations have already been taken place. On the other hand, visual detection aims to discover and locate patterns of interest, regardless of class interpretation, from a vast quantify of visual observations. For instance, for action recognition, a model is required to detect and segment instances of actions from a continuous observation of a visual scene. Detection in crowded scenes, such as detecting people fighting or falling in crowd, becomes challenging as objects of interest can be swamped by distracters and background clutters. To spot and recognise actions from a sea of background activities, the task of detection often poses a greater challenge than classification.

The problem of behaviour detection is further compounded when the behaviour to be detected is unknown a priori. A common aim of visual analysis of behaviour is to learn a model that is capable of detecting unseen abnormal behaviour patterns whilst recognising novel instances of known normal behaviour patterns. To that end, an anomaly is defined as an atypical and un-random behaviour pattern not represented by sufficient observations. However, in order to differentiate anomaly from trivial unseen instances or outright statistical outliers, one should consider that an anomaly satisfies a specificity constraint to known normal behaviours, i.e. true anomalies lie in the vicinity of known normal behaviours without being recognised as any.

1.1.3 Prediction and Association

An activity is usually formed by a series of object actions executed following certain temporal order at certain durations. Moreover, the ordering and durations of constituent actions can be highly variable and complex. To model such visual observations, a behaviour can be considered as a temporal process, or a time series function. An important feature of a model of temporal processes is to make prediction. To that end, behaviour prediction is concerned with detecting a future occurrence of a known behaviour based on visual observations so far. For instance, if the daily routine of a person's activities in a kitchen during breakfast time is well understood and modelled, the model can facilitate prediction of this person's next action when certain actions have been observed: the person could be expected to make coffee after finishing frying an egg. Behaviour prediction is particularly useful for explaining away partial observations, for instance, in a crowded scene when visual observation is discontinuous and heavily polluted, or for detecting and preventing likely harmful events before they take place.

Visual analysis of behaviour can assist object identification by providing contextual knowledge on how objects of interest should behave in addition to how they look. For instance, human gait describes the way people walk and can be a useful means to identify different individuals. Similarly, the way people perform different gestures may also reveal their identities. Behaviour analysis can help to determine when and where a visual identification match is most likely to be valid and relevant. For instance, in a crowded public place such as an airport arrival hall, it is infeasible to consider facial imagery identification for all the people all the time. A key to successful visual identification in such an environment is effective visual search. Behaviour analysis can assist in determining when and where objects of interest should be sought and matched against. Moreover, behaviour analysis can provide focus of attention for visual identification. Detecting people acting out of norm can activate identification with improved effectiveness and efficiency. Conversely, in order to derive a semantic interpretation of an object's behaviour, knowing what and who the object is can help. For instance, a train station staff's behaviour can be distinctively different from that of a normal passenger. Recognising a person as a member of staff in a public space can assist in interpreting correctly the behaviour of the person in question.

1.2 Opportunities

Automated visual analysis of behaviour provides some key building blocks towards an artificial intelligent vision system. To experiment with computational models of behaviour by constructing automatic recognition devices may help us with better understanding of how the human visual system bridges sensory mechanisms and semantic understanding. Behaviour analysis offers a great deal of attractive opportunities for application, despite that deploying automated visual analysis of behaviour to a realistic environment is still at its infancy. Here we outline some of the emerging applications for automated visual analysis of behaviour.

1.2.1 Visual Surveillance

There has been an accelerated expansion of closed-circuit television (CCTV) surveillance in recent years, largely in response to rising anxieties about crime and its threat to security and safety. Visual surveillance is to monitor the behaviour of people or other objects using visual sensors, typically CCTV cameras. Substantial amount of surveillance cameras have been deployed in public spaces, ranging from transport infrastructures, such as airports and underground stations, to shopping centres, sport arenas and residential streets, serving as a tool for crime reduction and risk management. Conventional video surveillance systems rely heavily on human operators to monitor activities and determine the actions to be taken upon occurrence of an incident, for example, tracking suspicious target from one camera to another camera, or alerting relevant agencies to areas of concern. Unfortunately, many actionable incidents are simply mis-detected in such a manual system due to inherent limitations from deploying solely human operators eyeballing CCTV screens. These limitations include: (1) excessive number of video screens to monitor, (2) boredom and tiredness due to prolonged monitoring, (3) lack of a priori and readily accessible knowledge on what to look for, and (4) distraction by additional operational responsibilities. As a result, surveillance footages are often used merely as passive records for post-event investigation. Mis-detection of important events can be perilous in critical surveillance tasks such as border control or airport surveillance. It has become an operational burden to screen and search exhaustively colossal amount of video data generated from growing number of cameras in public spaces. Automated computer vision systems for visual analysis of behaviour provide the potential for deploying never-tiring computers to perform routine video analysis and screening tasks, whilst assisting human operators to focus attention on more relevant threats, thus improving the efficiency and effectiveness of a surveillance system.

1.2.2 Video Indexing and Search

We are living in a digital age with huge amount of digital media, especially videos, being generated at every single moment in the forms of surveillance videos, online news footages, home videos, mobile videos, and broadcasting videos. However, once generated, very rarely they are watched. For instance, most visual data collected by surveillance systems are never watched. The only time when they are examined is when a certain incident or crime has occurred and a law enforcement organisation needs to perform a post-event analysis. Unless specific time of the incident is known, it is extremely difficult to search for an event such as someone throws a punch in front of a nightclub. For a person with a large home video collection, it is a tedious and time consuming task to indexing the videos so that they can be searched efficiently for footages of a certain type of actions or activities from years gone by. For film or TV video archive, it is also a very challenging task to search for a specific footage without text meta information, specific knowledge about the name of a subject, or the time of an event. What is missing and increasingly desired is the ability to visually search archives by what has happened, that is, automated visual search of object behaviours with categorisation.

1.2.3 Robotics and Healthcare

A key area for robotics research in recent years is to develop autonomous robots that can see and interact with people and objects, known as social robots (Breazeal 2002). Such a robot may provide a useful device in serving an aging society, as a companion and tireless personal assistant to elderly people or people with a disability. In order to interact with people, a robot must be able to understand the behaviour of the person who is interacting with. This ability can be based on gesture recognition, such as recognising waving and initialising a hand-shake, interpreting facial expression, and inferring intent by body posture. Earlier robotics research had focused more on static object recognition, manipulation and navigation through a stationary environment. More recently, there has been a shift towards developing robots capable of mimicking human behaviour and interacting with people. To that end, enabling a robot to perform automated visual analysis of human behaviour becomes essential. Related to the development of a social robot, personalised healthcare in an aging society has gained increasing prominence in recent years. To be able to collect, disseminate and make sense of sensory information from and to an elderly person in a timely fashion is the key. To that end, automated visual analysis of human behaviour can provide quantitative and routine assessment of a person's behavioural status needed for personalised illness detection and incident detection, e.g. a fall. Such sensor-based systems can reduce the cost of providing personalised health care, enabling elderly people to lead a more healthy and socially inclusive life style (Yang 2006).

1.2.4 Interaction, Animation and Computer Games

Increasingly more intelligent and user friendly human computer interaction (HCI) are needed for applications such as a game console that can recognise a player's gesture and intention using visual sensors, and a teleconferencing system that can control cameras according to the behaviour of participants. In such automated HCI systems using sensors, effective visual analysis of human behaviour is central to meaningful interaction and communication. Not surprisingly, animation for film production and gaming industries are also relying more on automated visual analysis of human behaviour for creating special effects and building visual avatars that can interact with players. By modelling human behaviour including gesture and facial expression, animations can be generated to create virtual characters, known as avatars, in films and for computer games. With players' behaviour recognised automatically, these avatars can also interact with players in real-time gaming.

1.3 Challenges

Understanding object behaviour from visual observation alone is challenging because it is intrinsically an ill-posed problem. This is equally true for both humans and computers. Visual interpretation of behaviour can be ambiguous and is subject to changing context. Visually identical behaviours may have different meanings depending on the environment in which activities are taken place. For instance, when a person is seen waving on a beach, is he greeting somebody? swatting an insect? or calling for help as his friend is drowning? In general, visual analysis of behaviour faces two fundamental challenges.

1.3.1 Complexity

Compared to object recognition in static images, an extra dimension of time needs be considered in modelling and explaining object behaviour. This makes the problem more complex. Let us consider human behaviour as an example. Human has an articulated body and the same category of body behaviours can be acted in different ways largely due to temporal variations, for example, waving fast versus slowly. This results in behaviours of identical semantics look visually different, known as large intra-class variation. On the other hand, behaviours of different semantic classes, such as jogging versus running, can be visually similar, known as small inter-class variation. Beyond single-object behaviour, a behaviour can be of multiple interacting objects characterised by their temporal ordering. In a more extreme case, a behaviour is defined in the context of a crowd where many people co-exist both spatially and temporally. In general, behaviours are defined in different spatial and temporal context.

1.3.2 Uncertainty

Based on visual information alone to describe object behaviour is inherently partial and incomplete. Unlike a human observer, when a computer is asked to interpret behaviour without other sources of information except imagery data, the problem is compounded by visual information only available in two-dimensional images of a three-dimensional space, lack of contextual knowledge, and in the presence of imaging noise.

Two-dimensional visual data give rise to visual occlusion on objects under observation. This renders not all behavioural information can be observed visually. For instance, for two people interacting with each other, depending on the camera angle, almost inevitably part of or the full body of a person is self-occluded. As a result, semantic interpretation of behaviour is made considerably harder when only partial information is available.

Behaviour interpretation is highly context dependent. However, contextual information is not always directly observable, nor necessarily always visual. For instance, on a motorway when there is a congestion, a driver often wishes to find out the cause of the congestion in order to estimate likely time delay, whether the congestion is due to an accident or road work ahead. However, that information is often unavailable in the driver's field of view, as it is likely to be located miles away. Taking another example, on a train platform, passengers start to leave the platform due to an announcement by the station staff that the train line is closed due to signal failure. This information is in audio form therefore not captured by visual observation on passenger behaviours. To interpret behaviour by visual information alone introduces additional uncertainty due to a lack of access to non-visual contextual knowledge.

Visual data are noisy, either due to sensor limitations or because of operational constraints. This problem is particularly acute for video based behaviour analysis when video resolution is often very low both spatially and temporally. For instance, a typical 24-hours 7-days video surveillance system in use today generates video footages with a frame-rate of less than three frames per second, and with heavy compression for saving storage space. Imaging noise degrades visual details available for analysis. This can further cause visual information processing to introduce additional error. For instance, if object trajectories are used for behaviour analysis, object tracking errors can increase significantly in low frame-rate and highly compressed video data.

1.4 The Approach

We set out the scope of this book by introducing the problem of visual analysis of behaviour. We have considered the core functions of behaviour analysis from a computational perspective, and outlined the opportunities and challenges for visual analysis of behaviour. In the remaining chapters of Part I, we first give an overview on different domains of visual analysis of behaviour to highlight the importance and relevance of understanding behaviour in context. This is followed by an introduction to some of the core computational and machine learning concepts used throughout the book. Following Part I, the book is organised into further three parts according to the type of behaviour and the level of complexity involved, ranging from facial expression, human gesture, single-object action, multiple object activity, crowd behaviour analysis, to distributed behaviour analysis.

Part II describes methods for modelling single-object behaviours including facial expression, gesture, and action. Different representations and modelling tools are considered and their strengths and weaknesses are discussed.

Part III is dedicated to group behaviour understanding. We consider models for exploring context to fulfil the task of behaviour profiling and abnormal behaviour detection. Different learning strategies are investigated, including supervised learning, unsupervised learning, semi-supervised learning, incremental and adaptive learning, weakly supervised learning, and active learning. These learning strategies are designed to address different aspects of a model learning problem in different observation scenarios according to the availability of visual data and human feedback.

Whilst Parts II and III consider behaviours observed from a single camera view, Part IV addresses the problem of understanding distributed behaviours from multiple observational viewpoints. An emphasis is specially placed on non-overlapping multi-camera views. In particular, we investigate the problems of behaviour correlation across camera views for camera topology estimation and global anomaly detection, and the association of people across non-overlapping camera views, known as re-identification.

References

Agre, P.E.: The dynamic structure of everyday life. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA (1989)

Barbur, J.L., Ruddock, K.H., Waterfield, V.A.: Human visual responses in the absence of the geniculo-calcarine projection. Brain **103**(4), 905–928 (1980)

Breazeal, C.L. (ed.): Desiging Sociable Robots. MIT Press, Cambridge (2002)

Buxton, H., Gong, S.: Visual surveillance in a dynamic and uncertain world. Artif. Intell. **78**(1–2), 431–459 (1995)

Gong, S., Ng, J., Sherrah, J.: On the semantics of visual behaviour, structured events and trajectories of human action. Image Vis. Comput. **20**(12), 873–888 (2002)

Gough, D.: The visual behaviour of infants in the first few weeks of life. Proc. R. Soc. Med. **55**(4), 308–310 (1962)

Gregory, R.L.: The Intelligent Eye. Weidenfeld and Nicolson, London (1970)

Harbluk, J.L., Noy, Y.I.: The impact of cognitive distraction on driver visual behaviour and vehicle control. Technical Report TP 13889 E, Road Safety Directorate and Motor Vehicle Regulation Directorate, Canadian Minister of Transport (2002)

Humphreys, G.W., Bruce, V.: Visual Cognition: Computational, Experimental and Neuropsychological Perspectives. Erlbaum, Hove (1989)

Ingle, D.J., Goodale, M.A., Mansfield, R.J.W. (eds.): Analysis of Visual Behaviour. MIT Press, Cambridge (1982)

Ltti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)

Marr, D.: Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information. Freeman, New York (1982)

Rollinson, D.: Organisational Behaviour and Analysis: An Integrated Approach. Prentice Hall, New York (2004)

Schiller, P.H., Koerner, F.: Discharge characteristics of single units in superior colliculus of the alert rhesus monkey. J. Neurophysiol. **34**(5), 920–935 (1971)

Sherman, M.D., Sherman, I.C.: The process of human behaviour. J. Ment. Sci. **76**, 337–338 (1930)

Simon, H.A.: A behavioral model of rational choice. Q. J. Econ. **69**(1), 99–118 (1955)

von Helmholtz, H.: The recent progress of the theory of vision. In: Popular Scientific Lectures. Dover, New York (1962)

Warrant, E.J.: Seeing in the dark: vision and visual behaviour in nocturnal bees and wasps. J. Exp. Biol. **211**, 1737–1746 (2008)

Weiskrantz, L.: Review lecture: behavioural analysis of the monkey's visual nervous system. Proc. R. Soc. **182**, 427–455 (1972)

Xiang, T., Gong, S.: Beyond tracking: modelling activity and understanding behaviour. Int. J. Comput. Vis. **67**(1), 21–51 (2006)

Yang, G.Z. (ed.): Body Sensor Networks. Springer, Berlin (2006)

Chapter 2
Behaviour in Context

Interpreting behaviour from object action and activity is inherently subject to the context of a visual environment within which action and activity take place. Context embodies not only the spatial and temporal setting, but also the intended functionality of object action and activity (Bar 2004; Bar and Aminoff 2003; Bar and Ullman 1993; Biederman et al. 1982; Palmer 1975; Schwartz et al. 2007). Humans employ visual context extensively for both effective object recognition and behaviour understanding. For instance, one recognises, often by inference, whether a hand-held object is a mobile phone or calculator by its relative position to other body parts such as closeness to the ears, even if they are visually similar and partially occluded by the hand. Similarly for behaviour recognition, the arrival of a bus in busy traffic is more likely to be inferred by looking at the passengers' behaviour at a bus stop. Computer vision research on visual analysis of behaviour embraces a wide range of studies on developing computational models and systems for interpreting behaviour in different context. Here we highlight some well established topics and emerging trends.

2.1 Facial Expression

Behaviour exhibited from a human face is predominantly in the form of facial expression. The ability to recognise the affective state of a person is indispensable and important for successful interpersonal social interaction. Affective arousal modulates all non-verbal communication cues such as facial expressions, body postures and movements. Facial expression is perhaps the most natural and efficient means for humans to communicate their emotions and intentions, as communication is primarily carried out face to face.

Automatic facial expression recognition has attracted much attention from behavioural scientists since the work of Darwin (1872). Suwa et al. (1978) made the first attempt to automatically analyse facial expressions from images. Much progress has been made in the last 30 years towards computer-based analysis and interpretation of facial expression (Bartlett et al. 2005; Chang et al. 2004; Cohen et al. 2003;

S. Gong, T. Xiang, *Visual Analysis of Behaviour*,
DOI 10.1007/978-0-85729-670-2_2, © Springer-Verlag London Limited 2011

Donato et al. 1999; Dornaika and Davoine 2005; Essa and Pentland 1997; Fasel and Luettin 2003; Hoey and Little 2004; Jia and Gong 2006, 2008; Kaliouby and Robinson 2004; Lee and Elgammal 2005; Lyons et al. 1999; Pantic and Patras 2006; Pantic and Rothkrantz 2000, 2004; Shan et al. 2009; Yacoob and Davis 1996; Yeasin et al. 2004; Yin et al. 2004; Zalewski and Gong 2005; Zhang and Ji 2005). A recent trend in human–computer interaction design also considers the desire for affective computing in order to bring about more human-like non-verbal communication skills to computer-based systems (Pantic and Rothkrantz 2003).

Facial expressions can be described at different levels. A widely used description is the Facial Action Coding System (FACS) (Ekman and Friesen 1978), which is a human-observer-based protocol developed to represent subtle changes in facial expressions. With FACS, facial expressions are decomposed into one or more action units. The aim to develop computer-based automatic AU recognition has attracted much attention in the last decade (Donato et al. 1999; Tian et al. 2001; Valstar and Pantic 2006; Zhang and Ji 2005). Psychophysical studies also indicate that basic emotions have corresponding universal facial expressions across all cultures (Ekman and Friesen 1976). This is reflected by most facial expression recognition models attempting to recognise a set of prototypic emotional expressions, including 'disgust', 'fear', 'joy', 'surprise', 'sadness' and 'anger' (Bartlett et al. 2005; Cohen et al. 2003; Lyons et al. 1999; Yeasin et al. 2004; Yin et al. 2004). Example images of different facial expressions are shown in Fig. 2.1.

Although a facial expression is a dynamic process with important information captured in motion (Bassili 1979), imagery features extracted from static face images are often used to represent a facial expression. These features include both geometric features (Valstar and Pantic 2006; Valstar et al. 2005) and appearance features (Bartlett et al. 2005; Donato et al. 1999; Lyons et al. 1999; Tian 2004; Zhang et al. 1998). The rigidity and structural composition of facial action components make facial expression a rather constrained type of behaviour, a raison d'être for extracting information from static images to discriminate different types of facial expression. Efforts have also been made to model facial expression as a dynamical process utilising explicitly spatio-temporal or image sequence information (Cohen et al. 2003; Shan et al. 2005; Zalewski and Gong 2005; Zhang and Ji 2005). These studies aim to explore a richer description of expression dynamics so to better interpret facial emotion and expression change.

2.2 Body Gesture

The ability to interpret human gestures constitutes an essential part of our perception. It reveals the intention, emotional state or even the identity of the people surrounding us and mediates our communication (Pavlovic et al. 1997). Such human activities can be characterised by hand motion patterns and modelled as trajectories in a spatio-temporal space. Figure 2.2 shows some examples of communicative gestures. Gesture recognition usually considers the problem of measuring the similarities between hand motion pattern characteristics, such as their trajectories, and