ELSEVIER

Image and Vision Computing 20 (2002) 825–826

image AND vision COMPUTING

www.elsevier.com/locate/imavis

Editorial

Understanding visual behaviour

The problems of modelling and understanding visual behaviours and their semantics are often regarded as computationally ill-defined. Many would argue that cognitive understanding cannot even adequately explain why we associate particular meanings with observed behaviours. In the words of Stevie Smith, we may be "Not Waving But Drowning". In general, it is certainly not true that meaningful interpretation of visual behaviour can rely on simple mappings from recognised patterns of motion to semantics. Understanding human activity, in particular, is complex as the same behaviour may have several different meanings depending upon the scene and task context in which it is performed. This ambiguity is exacerbated when several people are present in a scene. Is the observed behaviour purposeful, part of a communication and if so, how can we tell and can we extract the underlying meaning from visual data alone? Furthermore, do people behave differently in the presence of others and if so, how do we model to differentiate expected normal behaviours from those that have some degree of abnormality? In our example, a person waving on a beach could be just greeting somebody, swatting an insect or calling for help! Although these specific actions may be performed slightly differently according to their intended meaning, perhaps more urgent or intermittent motion, the question arises as to whether these differences can be sufficiently accurately measured. Further, whether differences in context such as somebody to wave to, an insect to swat, or rough water to be rescued from can be detected using visual information alone.

It is both significant and compelling then that one of the most noticeable developments in computer vision over the last decade has been the rapidly growing interest in the challenging problems of modelling and understanding human actions and behaviours captured in image sequences. This requires not only solving the problems of human detection, segmentation and tracking, motion trajectory analysis and classification, but also modelling of the semantics of both continuous motion patterns and discrete salient behavioural events. Behaviour interpretation often also requires real-time performance if it is to be correct in the relevant dynamic context. By real-time, however, it is not necessarily implied that all computation must be performed at full video frame-rate, as long as the interpretation of behaviour proceeds within some required time constraint.

Much progress has been made since Nagel's 1988 *Image and Vision Computing* article on 'From image sequences towards conceptual descriptions'. Most noticeably, statistical learning has become central to modelling actions and behaviours, which as one of the underlying themes is strongly reflected throughout this special issue. In general, the information processing involved in visual perception is both subject to noise and inherently ill-posed. This is especially true in modelling human activities captured visually. Segmenting and modelling human actions in a visual scene is far more difficult than rigid object movements due to the complex dynamics and deformable articulated structure of the human body. Whilst these visual phenomena are very difficult to model analytically, they can be effectively modelled probabilistically through statistical learning to capture reliable relationships in the visual data. Another important development in recent years is wider adoption of view-based representations. A view-based representation lends itself particularly well to learning from example images. Inevitably, such learning is to some extent context-dependent so example data must be selected to reflect typical as well as boundary conditions of the relevant context.

Three popular applications, visual surveillance, visually mediated interaction and visual animation, have undoubtedly played a significant role in recent years to highlight the issues above and bring about the necessary focus on understanding visual behaviour. This brings us to our special issue in which seven papers are selected to reflect the current trends, progress and issues on modelling and understanding human activity from visual behaviour.

A commonly adopted approach to modelling human actions for interpretation relies upon classification and labelling of motion trajectories of tracked human subjects of interest. Needless to say, the selection of moving subjects together with continuous and robust visual tracking by data association over time is essential. Li, Hilton and Illingworth present a method for real-time 3D tracking of human motion through multiple 2D camera views using a discrete relaxation algorithm. McAllister, McKenna and Ricketts describe a model for real-time tracking of a person's forearms and hands under self-occlusion and the interpretation of associated behavioural events using adaptive appearance models.

PII: S0262-8856(02)00092-6

Another approach is to treat the behavioural patterns observed from scene activity as trajectories in a high-dimensional feature space of correlated visual measurements. For example, the spatio-temporal patterns of a simple activity such as a person walking or running can be represented by the trajectory of a multivariate observation vector given by the position, speed and body boundary shape of the subject. Rittscher, Blake and Roberts present an approach to learning such a model for combined visual tracking and classification of human motion patterns for walking and running by learning the dynamics of a low-dimensional feature vector.

Statistical generative models not only attempt to reconstruct and describe faithfully the characteristics of some given (training) data, but more crucially aim to generalise and interpret any novel data using the models acquired from training. For example, stochastic processes or graph-based belief networks can be used to model the temporal structure of human activity in multi-dimensional feature space so that salient states or the nodes of a network are closely associated with behavioural events and the underlying causal semantics in a given scene. Johnson and Hogg describe a model for learning stochastic behaviour models of pedestrians and synthesising plausible trajectories using Gaussian mixtures. Gong, Ng and Sherrah present an approach for learning normal and abnormal behavioural events from pixel-energy-histories without object-centred segmentation and tracking, and for modelling the semantics of plausible human body motion configurations using Bayesian belief networks.

As we have mentioned above, the definition, interpretation and computability of normal and abnormal behaviours are largely context dependent. Automated learning of the *a priori* knowledge for associating normal behaviours with human activities in a given context is both challenging and extremely useful. One approach is to use statistical knowledge of a given scene to facilitate bootstrapping of learning and consequently recognition of abnormal behaviours. Makris and Ellis address the problem of learning models of common pedestrian paths in order to identify potentially suspicious behaviours in an outdoor scene.

Finally, the notion of visually mediated interaction in which a camera acts as an intelligent mediator between two remotely communicating parties intrinsically requires the understanding of human cognitive behaviour and its context based on captured images of the scene alone. To build such a system requires analysis of focus of attention, mapping of scene and camera views, and limited view-dependent reasoning. Howell and Buxton present an adaptive learning system for gesture and behaviour recognition in visually mediated interaction using time-delay radial basis function networks.

We are delighted to bring you this special issue which we hope not only serve as a sampling of recent progress but also highlight some of the challenges and open questions in automatic modelling and understanding of visual behaviour. We would like to thank all the authors for their contributions, the anonymous referees for their invaluable reviews, and Keith Baker, the editor-in-chief of *Image and Vision Computing,* for his support.

Shaogang Gong[a,*]
Hilary Buxton[b]
[a]*Department of Computer Science Queen Mary, University of London, London E1 4NS, UK*
E-mail address: sgg@dcs.qmul.ac.uk
[b]*School of Cognitive and Computing Sciences, University of Sussex, Falmer, Brighton BN1 9QH, UK*

* Corresponding author. Tel.: +44-207-975-5249; fax: +44-208-980-6533.