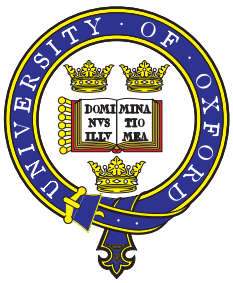
Automatic Causal Reasoning for Video Surveillance

Neil Robertson
Hertford College



Department of Engineering Science
University of Oxford

Hilary Term, 2006

This thesis is submitted to the Department of Engineering Science, University of Oxford, for the degree of Doctor of Philosophy. This thesis is entirely my own work, and, except where otherwise indicated, describes my own research.

Dedicated to the memory of my father.

Neil Robertson
Hertford College

Doctor of Philosophy
Hilary Term, 2006

Automatic Causal Reasoning for Video Surveillance

Abstract

This thesis is concerned with producing high-level descriptions and explanations of human activity in video from a single, static camera. The scenarios we focus on in this work are urban surveillance and sports video where the person is in the medium scale, around 150 pixels high. The final output is in the form of text descriptions which not only describe *what* is happening but also *explain* the interactions which take place.

In order to achieve this goal, some significant issues pertinent to action recognition and human behaviour estimation have been addressed. In particular, we have developed novel solutions for estimating where an imaged person is looking even when the face image is low-resolution. We have extended the Bayesian fusion techniques used to solve the gaze recognition problem to activity recognition in general. By computing non-static descriptors based on instantaneous target motion and combining them with position and velocity via an efficient non-parametric database search, we compute distributions over spatio-temporal actions. Probabilistic distributions over behaviour are further estimated from a set of Hidden Markov Models which encode stochastic sequences of actions. Automatic commentaries of most likely action sequences and/or higher-level behaviour at a human-readable level can be derived by computing the Maximum Likelihood or *Maximum a Posteriori* estimate at any time step, respectively. In the latter case we use domain knowledge as a smoothing prior to refine the estimates.

Finally, we draw these components together to achieve the main objective of this thesis: causal reasoning in video. Using an extensible, rule-based architecture we compute explanations of observed activity. The input to this reasoning process is the information obtained at the action/behaviour recognition stage, which represents an abstraction from the image data. The output of best explanations of global scene activity, particularly where interesting events have occurred, is thus achieved.

Acknowledgements

I am indebted to Ian Reid and Mike Brady for their inspiring and challenging supervision of this research. Ian and Mike, thanks for your encouragement, always positive attitude and belief that I would achieve this goal. Your desire to maintain the highest standards has resulted in me doing better, more interesting research than I could have imagined.

Thanks to Paul Newman and David Hogg, who examined this thesis, for your valuable time, insightful comments and helpful suggestions.

Thanks to my colleagues in Oxford and Malvern for interesting debate and much needed moments of light-relief. In particular to Alan Marrs, my industrial supervisor, thank you for encouraging me to apply for the Royal Commission for the Exhibition of 1851 Industrial Fellowship which has proved invaluable in the development of this work.

To Dave Hutber, thank you for giving me the opportunity to experience industrial research at DERA/QinetiQ and study for this doctorate. Your "bold and cunning plan" seems to have worked out in the end!

To my family, I thank you for your love and regular visits to The South. They are appreciated more than you know. Mum, you inspired in me a love of all things academic and I thank you for that gift.

Finally, and most significantly, to my beautiful wife: Kate, I cannot find words enough to thank you. Without your unfailing support, prayers and constant encouragement I would have given up long ago. Thank you for your amazing generosity which has enabled me to fulfill this dream.

Contents

1

Introduction

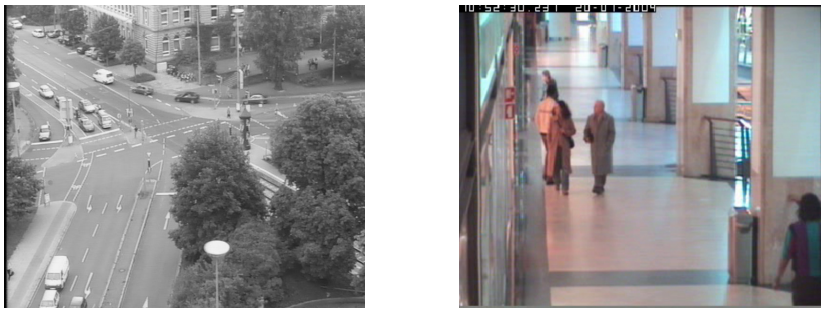


Figure 1.1: Typical surveillance scenarios in the civilian domain.

At the highest level, the scientific discipline of Computer Vision is concerned with enabling a computer to interpret the world, which is presented to it by one or more cameras, in a similar way to humans. Low-level vision techniques are necessary, and seem to occur at some point in the human visual process, but it is clear that the human visual system also operates at a higher "semantic" level. This enables people to make decisions based on their knowledge of the world, or their interpretation of it, and the evidence of their eyes.

Low-level vision, such as edge detection, does play a fundamental role in many Computer Vision systems, allowing scenes to be segmented into components which are indeed separate in reality, track objects and so on. But the question remains: *What principles of human vision can be modelled to enable a computer to "see"?*

Knowledge is a critical factor. Humans bring to bear a life's experience of seeing when presented with a new scene. And so, an individual's experience has an enormous impact on how he or she chooses to interpret a new situation. The world is infinitely complex and our knowledge of the world is finite so, as Kuipers points out, "The marvel is that we function quite well in spite of never fully understanding it" [86]. The simple application of knowledge is not the complete answer since, in the case of humans, there is the profound impact of *intelligence*. Human interpretation of visual data is ultimately in terms of causal relationships which are not found in the data alone, but are recognised by the mobilisation of prior knowledge, allied to an intelligent understanding of how the world operates.

Surveillance, which can be defined as the act of observing a person or group, involves the gathering of information useful to reasoning about intentions and behaviour extended over

time. "Intelligent Surveillance" is, therefore, the process by which we confer upon a machine the ability to produce human-level descriptions of observed activity. Provided, as Turing famously posited, that these computer-generated descriptions are hard to distinguish from those of a human expert, it could be claimed that Intelligent Surveillance has been achieved. Of course, it is a point for philosophical debate whether true intelligence can ever be achieved by a computer [113, 69].

There are, nonetheless, a number of issues which need to be addressed before Intelligent Surveillance will be achieved at all, whether in appearance or reality. This thesis addresses a subset of what we believe to be the current barriers to this goal, and considers what kind of framework is required for a general Intelligent Surveillance system.

The goal of this thesis is to develop a set of techniques to enable automatic causal reasoning about human activity as recorded in surveillance video.

1.1 Motivation

1.1.1 Military and civilian need

The motivation for this thesis stems primarily from a military need for better deployment of manpower resources. This is a pressing issue for the UK Ministry of Defence (MOD), which has, in part funded the research on which this thesis is based.

Automatic, or semi-automatic, video surveillance capability could enable Royal Air Force (RAF) Imagery Analysts (IAs) to focus on priority tasks without losing "situational awareness" i.e. without losing track of other events in the scene apart from the object on which they have been tasked to report. Currently, IAs work in pairs to avoid this problem arising. One IA focusses his attention solely on the target of interest while his/her partner scans the general scene looking for other interesting activity. An automatic tool which tracks all objects and reports on their activity at a high-level would enable this second analyst to work on another task since situational awareness is maintained by the system. This would prevent time, money and skilled resources being wasted and, as such, is a real driver for improved visual surveillance

techniques.

Such a system would also have many applications in the civilian domain due to the fact that the volume of information generated by a surveillance camera generally overloads all but the most expert of analysts. Included among these potential applications are: automatic surveillance of a shopping mall/homes, and Police surveillance of traffic. This high-level information can also be used for video meta-data[1] or text commentary[2].

1.1.2 Vision psychology

An intriguing set of results arises from a collection of experiments performed by psychologists in the early 1900s. These were first brought to attention in Michotte's book, *The perception of causality* [98], and in an article by Heider & Simmel, *An experimental study of apparent behaviour* [67] and have provoked considerable interest in the psychology literature.

The aim of the groundbreaking experiments in the works of Michotte and Heider & Simmel is broadly similar: to show that kinematics in simple moving displays gives rise to a perception which is not found in the events themselves or in the retinal projections of the events. Michotte's experiments generally involve two moving shapes which move with respect to one another in such a way as can be interpreted to interact. A sample of frames from a typical Michotte experiment is shown in Figure 1.2[3]. What Michotte, and Heider & Simmel, demonstrated is that such phenomena are important due to the fact that, although they are perceptual, they are "fairly fast, automatic, irresistable and highly stimulus-driven"[140]. Of particular interest to us is the fact that Scholl demonstrated

> the perception of causality is mediated by strict visual "rules". Beyond Michotte, ... these rules operate not only over discrete objects, but also over perceptual groups ... [29].

[1] For populating MPEG-7, for example.
[2] Of the type familiar to sports fans who use the BBC website.
[3] See `http://pantheon.yale.edu/~bs265/demos/` for this and other video demonstrations of Michotte's experiments.

Figure 1.2: Stills from the experiments of Michotte. Very simple motion gives rise to the perception of causality. In this example, the perception that the "object" on the left causes the "object" on the right to move is almost universal.

Perceptual causality, and its relevance to the work of this thesis, is discussed in more detail in chapter 2. In summary, however, we are motivated by this work in vision psychology in three important ways:

- Prior knowledge is necessary for interpreting dynamic, visual information,
- Motion, not appearance, is overwhelmingly the most important aspect of visual data for the perception of causality in the human visual system,
- A qualitative representation of motion is appropriate for human interpretation of causal interactions.

1.2 Objectives

"Surveillance video" is a term which appears many times in this thesis. What is it? In Figure 1.1 appear still frames which are typical of the type of footage one would consider surveillance video. The common factor between these images, from vastly different scenes, is essentially that the zoom level is sufficient to enable the object of interest to be clearly seen but also to allow the context of that object to be visible. This latter point is important because the actions of a person under surveillance can only be "suspicious" or "normal" or "interesting" within a specific context. It quickly becomes clear that the mobilisation of prior knowledge is a critical factor. We might, for example, detect someone standing still, but this could not be translated into the "intelligent" description "loitering" without knowledge of what normal behaviour is e.g. "standing still" is appropriate at a bus-stop, but far less so in an alleyway. A key factor of this thesis is therefore how best to incorporate human knowledge into the algorithms required for activity recognition.

Some of the components necessary for an effective video surveillance system have provided significant challenges for the scientific community. Tracking in video is one such problem. Simple, even naive, algorithms can achieve much. For example, foreground segmentation and "blob" identification can enable tracking throughout many image sequences. But the introduction of camera motion or occlusion within the video causes significant problems and sophisticated statistical techniques have been invented to solve what, for a grown adult, is a straightforward task.

We recognise that the tracking problem is not simple but it is clearly not the only important problem underpinning the search for a solution to intelligent visual surveillance. In this thesis we do not address tracking in any novel way since there are acceptable solutions to most tracking problems in the literature and since our major concerns are different. Therefore, we use a colour-based tracker which robustly tracks objects in scale and image-space providing the appearance does not alter dramatically throughout the sequence. More sophisticated trackers are available and may be required when the scenes become significantly more challenging than those we analyse in this thesis. The main weaknesses of the current tracking implementation are: (a) crowd scenes are excluded since individual people are only a few pixels in height and are often partially occluded; (b) it cannot handle significant occlusions coupled with target ambiguity such as two people who appear similar occluding one another meaning it is difficult to recover tracking using appearance alone.

Leaving tracking to one side then, what are the issues we identify and tackle? As we have stated, the overall goal is that we may be able to automatically reason about human activity in video. The key objectives become apparent when we consider what a human needs to know to *reason* about human activity:

- what is the object type?
- what is it doing?
- where is it doing it?
- what are the rules governing the actors in this scene?

- what are other agents doing?
- is there any connection between the agents' activities?

Given that intelligent visual surveillance is concerned with conferring upon a computer the ability to analyse video in a human-like manner, it is reasonable to assume that the computer requires access to the same information as a human would require to make a deduction. Moreover, previous work in the area of reasoning about visual data has been limited to static scenes with simple visual features. If this information about people can be made available to a reasoning process, then the techniques from AI could be utilised for the purpose of explaining complex, dynamic scenes.

The objective of this thesis is to develop new Computer Vision tools to obtain answers in response to the questions posed above. These tools operate on video where human activity is the dominant feature of interest, and where the imaged person is in medium/low resolution.

Our *thesis* can be concisely stated as:

> **In order for a computer system to effectively, and automatically, reason about human activity in surveillance video, low-level vision techniques must first abstract the information a human would require, from the video, to an intermediate, probabilistic and qualitative representation based on motion.**

The individual techniques we develop are outlined specifically in section 1.3 and 1.4. We first provide a summary of the main results of this thesis.

1.3 Achievements

Throughout this thesis, we have developed new algorithms and extended existing techniques for use in a video understanding context. Specifically, we make the following contributions:

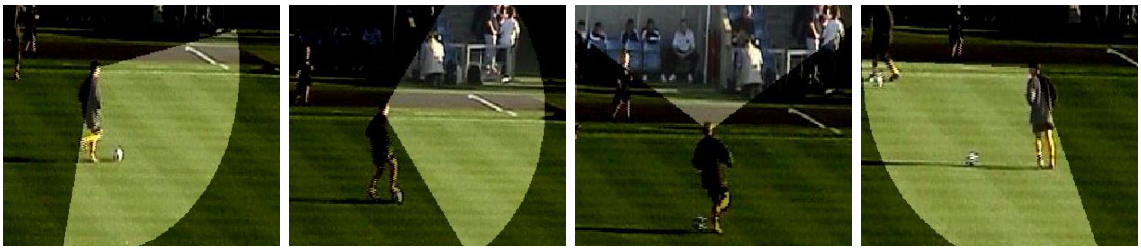


Figure 1.3: Gaze-direction is an important cue for intention. In chapter 3 we describe a novel method for estimating where someone is looking when the face of the imaged person is low-resolution.

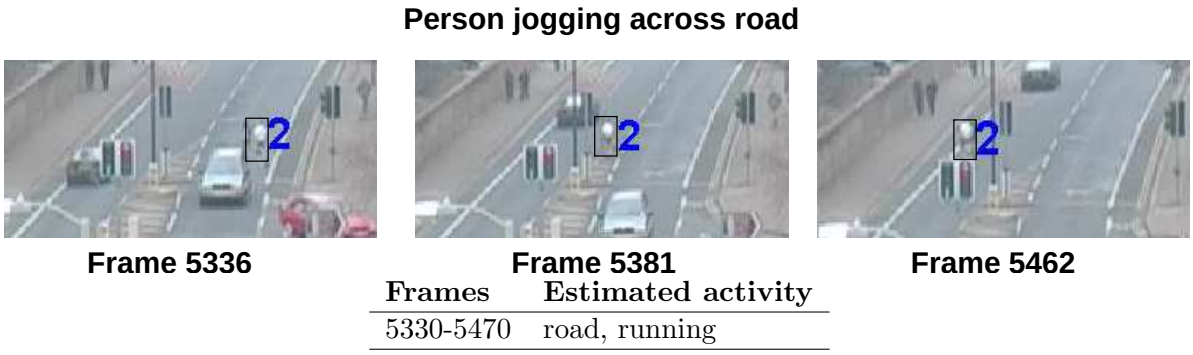


Figure 1.4: Action-recognition techniques, described in chapter 4, enable the estimation of distributions over the training data. The ML result is a commentary of activity, as shown here for a person jogging across the road.

- A novel method for gaze recognition in surveillance video where the face image is low-resolution i.e. typically in the range 20 to 40 pixels high. An example of the output of our algorithm is shown in Figure 1.3.

- A new, general Bayesian action-recognition framework. A set of hand-labelled exemplar databases which comprise the normal behaviour model for the scene under consideration is all we require to generate text descriptions of human action in video. Importantly, much less training data compared to the standard methods reported in the literature is needed. An example of this action-recognition method applied to surveillance footage is shown in Figure 1.4.

- A new, general method for encoding higher-level expert knowledge is achieved by exploiting the hierarchy of action and behaviour. Behaviour is represented as a sequence of actions, thus enabling behaviour to be described in a way that is common to the scene, not a specific viewpoint. This is made possible by abstracting the spatio-temporal actions

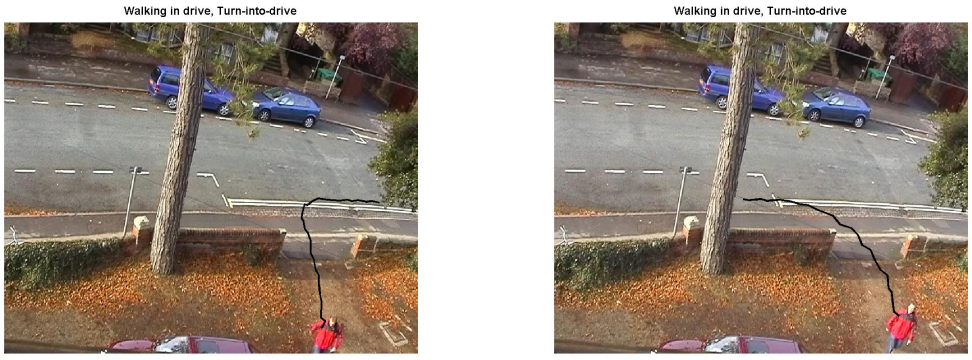


Figure 1.5: Encoding behaviour as a sequence of actions is more efficient than learning each example independently, as the work of chapter 5 demonstrates. In this example the same behaviour ("turning-into-drive") is acted out in two different ways (from the left and from the right), yet the same model correctly classifies both.

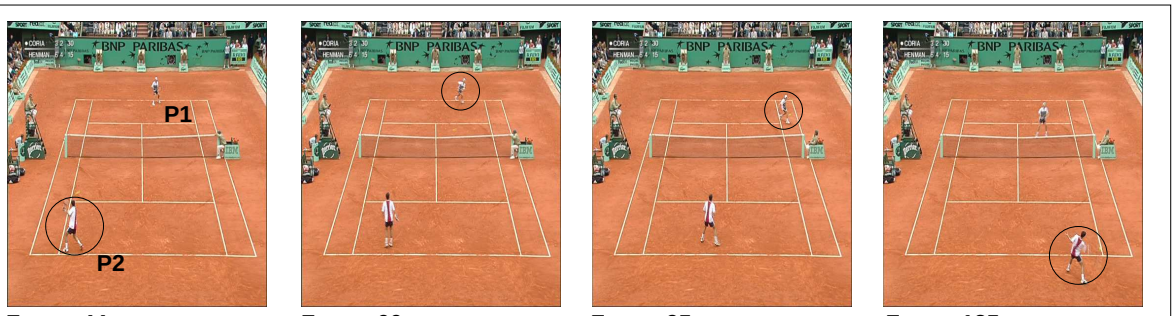
from the image pixel data to qualitative descriptions. An example of this technique in operation is shown in Figure 1.5.

- Explanations of interactions. By computing probabilistic distributions over gaze, spatio-temporal action and high-level behaviour for people in a video sequence, we provide the appropriate input to a rule-based architecture for causal reasoning. We achieve descriptions of interactions which not only say, in a human-readable form, *what* has happened (as a *maximum a posteriori* estimate) but can describe *why* it happened. An example of such an explanation is shown for tennis footage in Figure 1.6.

1.4 Approach

With the exception of chapter 6 where, for reasons of expediency, the Maximum Likelihood result has been used, we have adopted a Bayesian approach throughout this thesis which reflects the reasoning process which we, as experienced humans, employ. It is critically important to avoid committing to one interpretation of the activity in video because there is often incomplete information in any single view of the scene. The ability to hold one's view loosely in the face of a lack of evidence is an aspect of human reasoning which is particularly noteworthy.

As we have seen in section 1.1, it is necessary to incorporate prior knowledge in a principled way. One way to obtain prior knowledge is simply to observe a representative scene over a



Frame	Explanation
66-95	Player 1 forced to run-to-midcourt
96-135	Player 1 initiated run-to-net

Figure 1.6: Using the action/behaviour/gaze estimates enables reasoning about activity to be achieved in chapter 6. In this example, the explanation for a player running to the net is automatically extracted using a rule-based method. The rule-base is augmented and applied to the urban surveillance domain.

long period of time, an approach which is currently predominant in the literature. Taking this statistical approach, if one required to learn whether cars drive on the left or the right-hand side of the road, one could simply observe cars driving over a period of time and fit a model to the observations. This model then encodes the knowledge in which we are interested: the expected position of cars on a road in a certain country. However, the exemplar data must be mapped to an accurate text description, regardless of whether it is parameterised or not, in order to be useful for high-level reasoning. Otherwise, the best inference that can be made is that unusual activity is that which occurs less often. This may be true in a traffic situation where the "rules of engagement" are clearly defined and generally obeyed. But in a military scenario, such as urban combat or border control, it is not necessarily the case that frequency of occurrence is truly representative of normality. Therefore, it is appropriate when it is not guaranteed that large quantities of good training data are available, to enable human mark-up of the training data thus providing the correct normal model. Since the algorithms in this thesis have been developed with trained Imagery Analysts as the end users, we aim to utilise the extensive prior knowledge at their disposal[4]. This motivates the non-parametric recognition techniques (where the data itself is the model) which are used in much of this work. Due to the volume of this exemplar data, efficient search methods have been employed, which is novel in this context.

[4]This knowledge will, in future, be obtained by setting tasks as a result of which the IAs generate text reports.

The most basic information we can obtain about the activity of people in image sequences, using a video tracker, is position and velocity. This data alone is not of much use for reasoning about the general scene unless it is translated into a higher-level concept. Our hand-labelled exemplar databases provide this mapping between pixel data and human-readable labels. The database for each feature contains qualitative labels such that position and velocity, for example, become *place* and *direction*. As we discuss in more detail later in the thesis, there is much that can be achieved with this basic, qualitative information alone but a descriptive language at only one level of abstraction above the image data is an unnecessary and artificial limitation. We therefore consider spatio-temporal action, using Bayesian fusion, and behavioural information, which is represented by stochastic sequences of spatio-temporal actions.

First, however, we develop a novel method for determining in which direction a person is looking in low-resolution video. This technique is based on skin detection and we use body-direction as contextual information to compute distributions over potential gaze angles. Then, inspired by recent work in human action recognition, we extract descriptions of motion such as "walking", "running" etc. using instantaneous motion descriptors based on optic flow. By fusing the estimated probability distributions over all possible places and speeds with this instantaneous target-centred action description we derive higher-level descriptions of spatio-temporal motion e.g. "walking North on the West pavement". Expert knowledge is then used to create behaviour models which encode transitions between spatio-temporal actions.

We demonstrate that this action/behaviour information can be used to generate an automatic commentary on human activity in video sequences. We further show that smoothing priors based on expert knowledge improves this commentary. The complete description of activity (including gaze information) is finally passed to a rule-based reasoning system which derives explanations of interactions between agents in complex, real-world scenarios. All of these techniques are demonstrated extensively on urban surveillance and sports video.

1.5 Roadmap

While contributing towards the overall goal, each chapter in this thesis is, to some degree, self-contained and describes a component of work which is of interest in its own right. As such the chapters have their own summary and conclusion. To aid the reader and maintain continuity, we review the relevant literature which has not been covered in the main literature review at the start of each chapter.

Chapter 2 sets the scene by reviewing the strengths and weaknesses of the scientific state-of-the-art as reported in the peer-reviewed literature. This chapter puts the contributions of this thesis in the context of the latest work in visual surveillance.

Chapter 3 introduces the Computer Vision methods applied to the task of estimating where someone is looking when the imaged face is low-resolution. We demonstrate the use of temporal and contextual information for refining probabilistic estimates from static imagery.

Chapter 4 deals with recognising patterns of motion which correspond to human activity and estimating the spatio-temporal action of one person in video. This chapter addresses the problem of generating a probabilistic interpretation of the observed action via non-parametric sampling from an exemplar database.

Chapter 5 demonstrates the efficacy of representing behaviour as a stochastic sequence of actions, chiefly exploring the use of Hidden Markov Models in this context.

Chapter 6 draws the work of the previous chapters together in order to demonstrate causal reasoning. The text descriptions of activity in terms of spatio-temporal actions, extended behaviour and gaze-direction is used as input to a rule-based engine. Final examples are shown in this chapter on surveillance-style video in an urban setting and on sports footage.

Chapter 7 summarises the main achievements of this work and discusses avenues of future research based on our algorithms and results.

2

Related Work

In this chapter we provide an overview of the significant areas in relation to this thesis: vision psychology, Bayesian methods, modelling human behaviour in video and visual surveillance.

2.1 Vision psychology

The experiments of Michotte and Heider & Simmel are interesting and relevant to researchers not only in psychology and psychophysics but also in computer vision more generally because it has been demonstrated that very simple visual motions give rise to surprisingly high-level percepts. Moreover, the vast majority of viewers construct the *same* interpretation. The natural question to ask in light of these remarkable and compelling demonstrations is, is it possible to exploit the aspect of the motion causing these powerful perceptions to interpret visual scenes in causal terms in a computer vision system?

First, Michotte produced a series of simple moving displays which are readily interpreted as causal relations. An example of which can be described as follows (and is seen in Figure 1.2):

> Two small circles are sitting in a line, separated. The first circle A moves in a straight line until it reaches the second circle B, at which point A stops moving and B starts moving along the same trajectory.

In an objective sense nothing more is happening in the scene. The typical human interpretation is however that the motion of A *causes* the motion of B. Psychologists draw a distinction between the *inference* and the *perception* of causality. For our purposes, it is the fact that the phenomena arise at all that is of interest. Moreover, it is critical that Michotte pared the experimental stimuli down to the slightest of visual cues, namely the precise image motion, which produce these effects. This was done to demonstrate the importance of time: the perception of causality is reported to be much less powerful if there is a delay in both shapes moving after they appear to touch and disappears if the temporal gap is large enough.

It is conjectured that one reason for such a dramatic effect is that certain assumptions appear to be hard-wired into the human visual system e.g. the heuristic assumption that objects are rigid which is in play when extracting structure from motion [155]. Scholl believes this is what Michotte, Heider and Simmel have done for causality and animacy: used simple schematic displays which satisfy the assumptions in the most minimal way possible [140]. In fact this

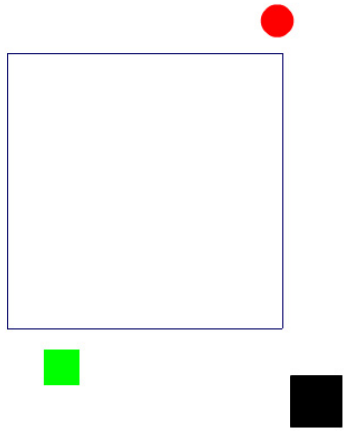


Figure 2.1: Still from Heider and Simmel film. In the animation the coloured objects move in such a way that not only is causality almost universally perceived but, for many viewers, emotional states are attributed to the shapes.

seems to have been done with such success that the slightest objective change in the displays can cause the perceptual effects to disappear completely. It has been demonstrated that the *kinematic* nature is the most important aspect for the perception of causality. So in Figure 1.2 the masses of the "balls" (or other physical properties of the system) are not required to be known in order that the causal relationship be perceived.

Michotte's claims are strong, and, unsurprisingly, some researchers contest whether the perceptual effect is as immediate and irresistable as suggested. White stressed that this should not be allowed to distract from the existence of the phenomena:

> The remarkable thing . . . is that causal processing is sufficiently irresistable to occur at all with such imperfect stimuli [165].

We might also add that it is remarkable that the "causal processing" occur so universally.

Heider and Simmel's work justifies the second half of the claim that not only perceptual causality but animacy can be produced by sparse visual cues. Not only is the perception of an object being alive in terms of being able to *cause* actions possible, but also there arises the perception of goals and emotional states e.g. "*wanting* to get to the red block" (a still from the famous Heider and Simmel sequence are shown in Figure 2.1). The film which Heider and Simmel created shows three geometric figures, a large square, a small square and a small circle, moving

near a rectangle. In their report of experiments, observers attributed personalities to the objects, such as shyness and bullying in addition to emotions, such as anger and frustration, regardless of the instructions they were given. Later work which suggested that the results are sensitive to context and that for the experiments to succeed the subjects must be primed with emotional information is inconclusive [143]. In defence of the conclusions of Michotte [99], and Heider and Simmel, further studies have shown that infants understand that inanimate objects cannot act on one another from a distance and so must have derived an impression of animacy from the quality of the objects' motions [138].

Notwithstanding the intrinsic interest and fascination of this psychological research, the question must be posed, what specifically is the relevance of this research to dynamic scene understanding and in particular the detection of interesting or suspicious behaviour? The answer is *motion cues*. Michotte suggested that simple motion cues are the foundation for social perception in general:

> In ordinary life, the specifying factors - gestures, facial expressions, speech - are innumerable and can be differentiated by an infinity of nuances. But they are all additional refinements compared with the key factors, which are simple kinetic structures [18].

In other words, it seems that psychophysics and perceptual psychology have provided a wealth of evidence illustrating how important, indispensable and fundamental motion is to humans as they try to make sense of the world they see. Given that computer vision is, in essence, attempting to confer on computers the power to see the world around them in terms of being able to understand and make intelligent decisions as humans can, it is only sensible to attempt to use those minimal cues which humans require.

A compelling result is that the kinematics, rather than the dynamics, of the scene is what produces the perceptual effects of causality and animacy. This distinction is important as it takes us from a quantitative position to a qualitative interpretation, where precise definitions of mass, velocity etc. are not as important as large-scale changes in direction, relative velocity

etc.

In addition, it has been suggested that humans employ a low-level "intuitive physics" which correlates geometrical properties of distance and size with physical properties such as mass, velocity and acceleration [43]. Cooper and Munger [32] believe this is done by internalised kinematic principles. It is known that very sparse clues can provide enough information for people to identify walking, for example even though the only visible information provided is small dots placed on an invisible human figure. The static representation cannot be recognised [77] but the motion of the dots allows subjects not only to identify the human walking but also to make fine distinctions regarding gait e.g. identify walking, running, limping etc. The conclusion is that non-rigid motion gives a sense of animacy.

2.2 Bayesian methods for data modelling

A visual process which surveys the world using a camera will generate a vast amount of data. The task of a surveillance system is to make sense of the data, detecting patterns which can be identified as certain modes of behaviour. Methods which model the data are useful so long as a less complex representation of the data than the data itself is offered. The complexity of the model is generally identified by the number of model parameters. This however introduces a conflict since the model fit is improved monotonically with increases in complexity [128], meaning that a model which is as complex as the data will fit perfectly. In order to achieve a balance between these measures, a number of methods are proposed in the literature with regard to graphical models.

Graphical models are used to compactly illustrate joint probability distributions. Bayesian Networks (Bayes Net) have become a common tool in statistics and probability. Initially introduced by Pearl [110], they can be interpreted as encoding causal relations. In fact, it is to this very interpretation that Pearl attributes the popularity of the Bayes Net representation.

> The interpretation of direct acyclic graphs [Bayes Nets] (DAG) as carriers of independence assumptions does not necessarily imply causation; in fact, it will be

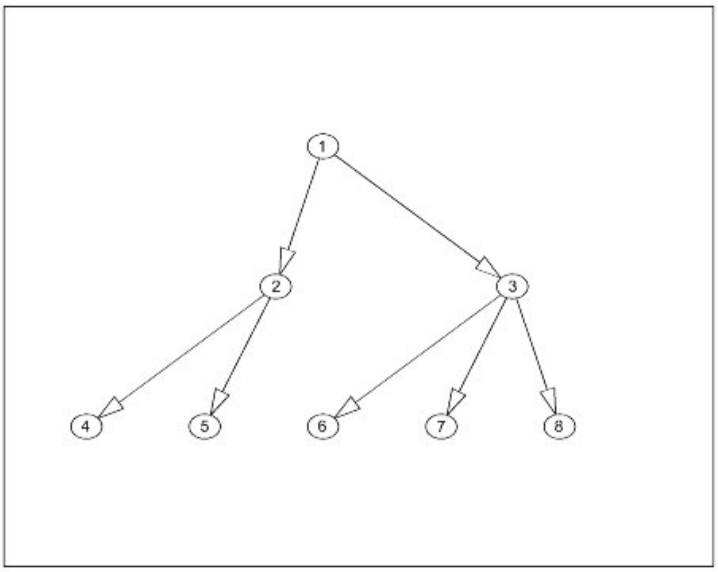


Figure 2.2: A simple Bayesian network graphically illustrates conditional independence. The topology of this graph tells us that if the value at node (3) is known, knowing the value of node (1) provides no more information at nodes (6),(7) or (8).

> valid for any set of recursive independencies along any ordering of the variables, not necessarily causal or chronological. However the ubiquity of DAG models in statistical and AI applications stems (often unwittingly) primarily from their causal interpretation - that is a system of processes, one per family, that could account for the generation of observed data [111].

The nodes of the graph represent random variables, and, in the case of a directed graphical model, the notions of interdependence of the variables takes into account the direction of the arcs joining the nodes. Although directed models have a more complicated notion of independence than undirected models, they do have several advantages. The most important is that one can interpret an arc from A to B probabilistically. Further when the concept of time is introduced this idea can be extended to Dynamic Bayesian Networks (DBN) where a Bayes Net is "rolled out" in time, a popular example of which is the Hidden Markov Model (HMM).

Bayes Nets are directed graphs representing stochastic processes. When specifying the Bayes Net, the topology is generally predefined and it is this structure of the graphical model that captures the the relationship of the states and the possible transitions between states in a conceptual manner. A graphical model specifies a complete joint probability distribution over

all the variables and the conditional independence relations can be immediately identified, as illustrated in Figure 2.2. If the graphical model parameters and structure can be learned from the data using, for example, Expectation Maximisation (see appendix C and [60]) then the task of fitting a model to the data can be accomplished without supervision of the learning process.

Note that despite the inclusion of "Bayesian" in Bayes Net, this does not commit the user to the Bayesian methods for *learning* which will be outlined here. Rather it indicates that Bayes' rule is used for probabilistic inference. Bayes' rule essentially provides a mathematical rule for how one should change one's beliefs in the light of new evidence. The probability that event R was caused by e is given by:

$$P(R|e) = \frac{P(e|R)P(R)}{P(e)} \tag{2.1}$$

The *posterior* is $P(R|e)$, the *conditional likelihood* is $P(e|R)$, the *prior* is $P(R)$ and the *evidence* is $P(e)$.

Roberts *et al.* argue in [128] that a Bayesian approach may be regarded as estimating the uncertainty of the model as a whole, given the data, and also estimating the uncertainty in the parameters. The uncertainty of the model decreases with the number of parameters, while the uncertainty in the parameters increases as more parameters are estimated. Therefore Bayesian modelling involves finding a balance between the two measures. The results of Bayesian modelling are compared with other popular model selection criteria including Minimum Description Length (MDL), Minimum Message Length (MML), Evidence Density, Partition Coefficient and the evidence is that model selection based on information theory i.e. MDL, MML and Bayesian methods in general outperform more heuristic methods [128].

2.2.1 Bayesian Networks

The example Bayesian Network in Figure 2.3 is trivially simple but illustrates in a straightforward way how Bayes Nets compactly represent conditional probability relations. The probabil-

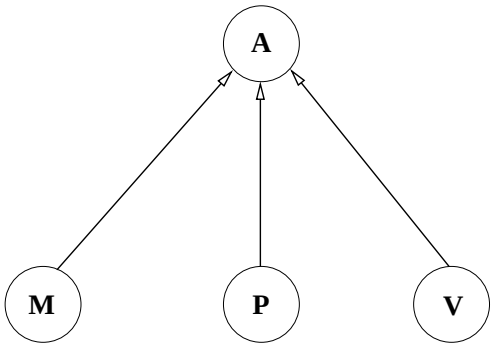


Figure 2.3: This is an even simpler example of a Bayesian network but it is useful for demonstrating the utility for representing mathematical relations compactly, as shown in the text.

ity $p(A)$ can be computed as the marginal probability at the node A:

$$p(A) = \sum_{P} \sum_{V} \sum_{M} p(A, P, V, M) \tag{2.2}$$

(Note that the choice of A,P,V,M as variable relates to action, position, velocity and motion, respectively which are described fully later in this thesis.) We can express the joint distribution of the Bayes Net in terms of the conditional relationship between the variables using the chain rule of probability

$$p(A, P, V, M) = p(A)p(P|A)p(V|A, P)p(M|A, P, V) \tag{2.3}$$

which can be simplified using the conditional independence of the leaf nodes to

$$p(A, P, V, M) = p(A)p(P|A)p(V|A)p(M|A) \tag{2.4}$$

Now, given the values that are specified in the conditional probability table for node A it is possible to compute the marginal distribution $p(A)$ since, for any given data D, (composed of

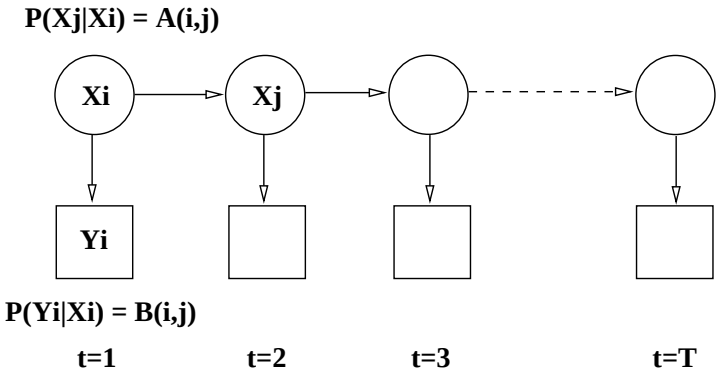


Figure 2.4: A Hidden Markov Model is a specific example of a Dynamic Bayesian Network and is completely defined by a set if transition probabilities, observation probabilities and (initial) state priors. The hidden variables are shown as circles and the observations as boxes. The states (X) transition probabilities are defined by A, the observation probabilities by B.

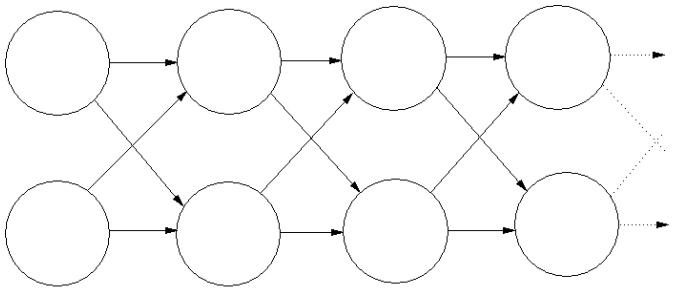


Figure 2.5: A Coupled HMM is just one of a number of variants on the standard HMM shown in Figure 2.4. The hidden states are coupled in state-space and the inference therefore includes a joint distribution over hidden states (see [123]).

$P, V, M)$ using Bayes rule:

$$p(A|D) = \frac{p(D|A)p(A)}{p(D)} \tag{2.5}$$

The priors on the possible values of A and D are predefined (or taken to be uniform) and the conditional probabilities are specified in the Conditional Probability Table (CPT) for this Bayesian Network.

2.2.2 The Hidden Markov Model

The Hidden Markov Model is an example of a *Dynamic* Bayesian Network (DBN). A DBN has the properties of a static Bayes Net of the type shown in Figures 2.2 and 2.3 but the time slices are typically shown chronologically left-to-right, as shown in Figure 2.4. The topology shown in Figure 2.4 is the most basic, standard HMM. HMMs have topology tailored to specific problems e.g. Coupled HMMs, as shown in Figure 2.5 [123]. The Markov assumption is that the future is conditioned only on the present, meaning there is no "memory" in the state sequences, as shown in Figure 2.4.

An HMM is a stochastic state-space model analogous to a Kalman Filter which can change state and is defined by the following parameters (collectively Θ):

$$\Pi = (\pi_i) \tag{2.6}$$

$$A = (a_{ij}) \tag{2.7}$$

$$B = (b_{ij}) \tag{2.8}$$

where Π is the initial state probabilities, A is the (hidden) state transition probabilities and B is the observation (of the hidden state) probabilities. These matrices typically are static i.e. the values do not change over time, which is one of the most unrealistic properties of the Markov model, in practice. They can take discrete or continuous (e.g. Gaussian or multi-variate Gaussian) distributions. A limitation of the HMM is the assumption that successive observations are independent and therefore the probability of a sequence of observations $p(O_1, O_2, \ldots, O_T)$ can be written as a product of individual observations:

$$p(O_1, O_2, \ldots, O_T) = \prod_{t=1}^{T} p(O_t) \tag{2.9}$$

Notwithstanding these limitations, the HMM has found practical application in a wide number of areas, particularly speech recognition [120], mainly due to the discovery of efficient algorithms for the calculations associated with the HMM model, which are discussed below.

A Hidden Markov Model has three main uses:

- Evaluation
- Decoding
- Learning

Evaluation The behaviour of an animal, for example, differs according to the season. We may have a set of HMMs which encode the expected behaviour for each season, summer, winter etc. Given an observation sequence of the animal behaviour $O = O_1, O_2, \ldots, O_T$ and a model e.g. summer, $\Theta = (A, B, \Pi)$ the probability of the observation sequence given the model is computed using the *forwards algorithm* [6, 7] (which is derived in appendix D). In speech recognition, for example, this problem occurs when someone speaks since an observation sequence is generated. A bank of word models is predefined and the most-likely model is selected to explain the observations, in this case the spoken words.

Decoding This problem can be stated as that of finding the most probable sequence of hidden states given some model and a set of observations i.e. given $O = O_1, O_2, \ldots, O_T$ and $\Theta = (A, B, \pi)$ how do we choose a corresponding state sequence $Q = q_1, q_2, \ldots, q_T$ which is optimal? The hidden states are often of interest because they represent something not observable (although there is no guarantee that the hidden states have any physical meaning). For example, it may be that only the animal behaviour (perhaps remotely) can be observed. From the observations weather state - the hidden state - perhaps temperature, can be inferred. The Viterbi algorithm is used to determine the most probable state sequence given a sequence of observations and a HMM (see appendix D).

Learning How do we adjust the model parameters $\Theta = (A, B, \Pi)$ to maximise $p(O|\Theta)$? This is the most difficult problem. The forward-backward algorithm is useful when the matrices A and

B are not directly measurable[1]. Expectation Maximisation can be used but is only effective when a good initial estimate of the model parameters can be provided.

2.2.3 Learning graphical models

A standard method used to learn the parameters of a graphical model is Expectation Maximisation (EM) (used later in this thesis and outlined in appendix C). When applied to a dataset for the purposes of training, EM returns an estimate of a single optimal value for the parameters within a fixed graph structure. The limitations of EM include a tendency to over-fit the data and a preference for complex models. The latter arises because more complex models have an increased number of parameters and so provide a better "explanation" of the data. But EM cannot optimise model structure [3, 107]. Roberts et al. [128] offer a solution (in principle) which is not subject to these limitations. For each model, the posterior probability of the model given the data is calculated. Predictions for the training data are then computed by averaging the predictions of all the individual models, thus avoiding over-fitting. Complex models are classified as more unlikely by being given a low posterior probability and therefore the choice of structures can be considered optimal.

Even for basic models, the computational effort often results in the Bayesian method becoming intractable [11]. Approximation methods include Markov Chain Monte Carlo (MCMC) and large sample methods.

Recently an additional method has been introduced: Variational Bayes [3]. This framework facilitates analytical calculations of posterior distributions over the hidden variables, parameters and structure. Moreover, they are computed via an iterative algorithm closely related to Expectation Maximisation (EM). One perceived advantage is that the MDL criterion emerges as a limiting case. Variational Bayes has been shown to outperform MDL for 1-dimensional Gaussian mixture models [112]

[1] Although we do not define this algorithm explicitly but it can readily be inferred from the forwards and Viterbi algorithms.

2.3 Detection and interpretation of human activities in video

Figure 2.6: Traffic behaviour prediction in the work of Brand and Kettnaker (from [22]). *Right*, predicted spatial positions of the car are shown as ellipses.

There is a wealth of literature on the variety of problems associated with the detection of human behaviour in video. Researchers have tackled issues such as tracking in clutter [157], face recognition, expression analysis [45] and human motion capture [61]. While recognising the importance of this work and the basis it has provided for subsequent robust interpretation of detected activity in video, in this section we explicitly consider only prior work which has focussed on the learning and modelling of human behaviour since that is what is directly relevant to the topic of the thesis.

It is worth noting that, while we have explicitly ignored person detection to automatically initiate a tracking process, this problem has been the focus of a number of recent innovations in the literature. In particular Viola *et al.* have extended their static object detection technique to achieve robust detection of pedestrians in video where people are around 100 pixels high [158]. Illustrative results from their work are shown in Figure 2.7.

Cutler and Davis pushed the detection problem to a remarkable limit by tackling the problem of detecting people from Unmanned Air Vehicle footage. The resolution of a person is very low (see Figure 2.8), but using the periodic motion of the walk, detection is possible, as shown in [36].

As highlighted by Oliver *et al.* [107] there has, over the last decade or so, been an increasing interest in the problem of analysing human behaviour in video [16, 25, 47]. Currently a system developed for such a task would generally be composed of two major components:

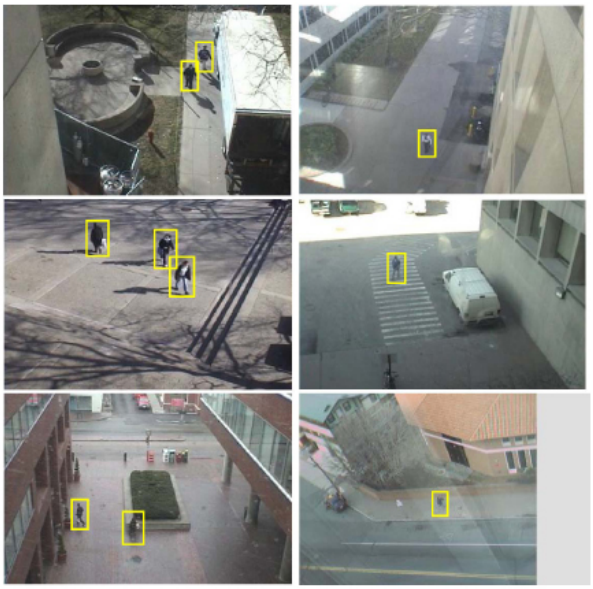


Figure 2.7: Viola *et al.* demonstrated robust detection of pedestrians in surveillance video [158]. This can be used as an initialisation to a tracker. We explicitly use a semi-automatic process, however.

Figure 2.8: From Cutler and Davis [36]. Detection of people can be achieved in very low resolution video.

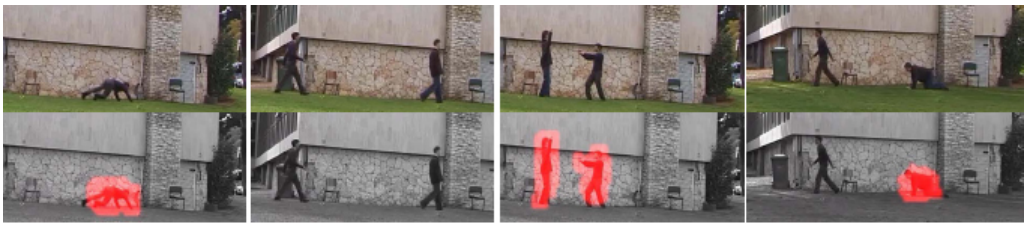


Figure 2.9: From Boiman and Irani [17]. Unusual activity is highlighted in this sequence (red patches, bottom row), by searching for pieces which resemble patches in the input frame in the training database. In this example, the training data is footage of someone walking backwards and forwards in the same scene from the same camera viewpoint.

1. low-level detection or segmentation of moving objects,

2. higher-level interpretation of tracks to classify the behaviours taking place in the scene.

A particular application which has been identified as an area on which little research effort has been expended is that of human behaviours which extend in time [108]. This task necessitates a combination of computer vision and artificial intelligence. Computer vision because accurate, robust detection and tracking of objects is required; artificial intelligence because knowledge representation and mobilisation are needed to interpret perceived actions and potentially dangerous behaviours (due to the reliance on prior models of behaviour in humans). These components are required, of course, to be integrated in some fashion. Notwithstanding this clear need, a limitation of the currently reported machine learning approaches based on example or data driven methods is that for rare or abnormal actions there is generally less abundant data for training, as one would expect.

A piece of work which spans a number of technical areas with which this thesis is concerned has been reported by Boiman and Irani [17]. The work addresses the problem of detecting "irregularities" in video, where "irregular" is defined solely by the context in which the video takes place. Essentially, if parts of the video (defined as the "query"), that is patches within frames extended in time, can be composed from other patches within the sequence (defined as the "database") then the patch is considered normal. Otherwise the patch is highlighted as an irregularity, as shown in Figure 2.9.

Boiman and Irani use a probabilistic graphical model for inferring the likelihood of observed activity given the video patch database. Although they claim to be using "no prior knowledge",

Figure 2.10: From Boiman and Irani [17]. In this example, there is no prior knowledge about what is normal behaviour, and no corresponding database of video patches representing normal behaviour. The activity highlighted in red is detected as abnormal by comparing patches within the video itself, searching for support. Patches with the least support are said to be "irrregular".

in reality it can be argued that, for sensibly detecting irregularity, the training data must implicitly be a model of normality, otherwise the predicted "suspicious" behaviour may not be suspicious at all, merely infrequent. The spatio-temporal patches which are size 7×7×4 are computed over a pyramid of spatial and temporal scales. One important issue which is discussed by Boiman and Iran is the time it takes to search this database. The search is posed as an inference problem, but, by progressive elimination, the complexity is, at worst, $\sim O(N)$, where N is the number of patches in the database, quoted as $N = 100,000$.

The final, and perhaps most interesting results, show, as initially claimed, the detection of irregularity in video with no prior information and no database. This example is shown in Figure 2.10. However, in a surveillance context, which seems to be the application the authors have in mind, it is not clear that this approach could yield robust anomaly detection due to the wide variety of appearances the same, innocuous, activity can take. The fact that the activity on which this method is demonstrated is so very structured (i.e. people waving hands in the air) may limit the range of domains in which the work can be applied.

2.3.1 Parametric methods

Successful examples of using a Bayesian [87] method are available in current literature [107, 22]. In particular, potential problems such as learning novel behaviours from as few as one example have been addressed using a Bayesian method incorporating graphical models [24] such as Hidden Markov Models (HMM)[120]. A large number of papers detail research using HMMs for post-visual processing and event classification. In most cases each HMM is trained on a number of examples of a given event and novel events are classified using likelihood ratios - a

standard Bayesian *maximum a posteriori* (MAP) approach. The topology of the HMM, based on a graphical model, is a crucial factor and is the subject of intense research outside of the standard vision literature [21].

The applications in these recent papers include areas of specific interest as far as detection of suspicious behaviour is concerned, specifically analysis of human interactions over long and short time scales and traffic monitoring. Despite differences in the details of constructing the topology of the relative learning models in terms of the HMM, it has been adequately demonstrated that classification and analysis is possible by means of unsupervised learning [22].

To highlight one example, Stauffer and Grimson aimed to:

> develop a visual monitoring system that passively observes moving objects in a site and learns patterns of activity from those observations [148].

As expected, the tracking of objects through video sequences is a vital component of this system and it is accepted that a tracking system for use in surveillance should be robust. In other words, the successful track initialisation (identification of when and where in image space to begin a new track) and track maintainence (the ability to identify the tracked object from the previous frame in the current frame) does not depend on clever placement of cameras or choosing carefully the illumination conditions. So for research with successful outcomes in the area of surveillance it is assumed tracking is enabled, robustly dealing with changes in scene illumination in particular.

In fact, many systems have been created to aid urban surveillance, most based on the notion of trajectories alone. For example Grimson *et al.* [63] report an entirely automated system for visual surveillance and monitoring of an urban site using agent trajectories. The same is true in the work of Buxton (who has been prominent in the use of Bayesian networks for visual surveillance) [26], Morellas *et al.* [101] and Makris [92]. Johnson and Hogg's work [78] is another example where trajectory information is only considered.

Galata *et al.* [52] address more specifically the problem of learning behavioural patterns. The

authors draw an important distinction between their work and prior work in the field by stating that the use of Variable Length Markov Models (VLMM) allows the structure of underlying training data to be automatically inferred. This is an important point as the work of Brand and Kettnaker [22] requires one model to cover the entire behaviour space and that in turn is in contrast to most of the previous techniques which have been discussed which have separate, predefined models. To illustrate the main idea behind the VLMM, assume there is a string of tokens, s, used as a memory to predict the next token, t' according to an estimate $\hat{P}(t'|s)$ of the true probability $P(t'|s)$. If the probability $\hat{P}(t'|ts)$ that predicts the next token is significantly better then the the longer memory ts is a better predictor than s. From [52], there is a measure to determine how much additional information is gathered using the longer memory:

$$\Delta H(ts, s) = \hat{P}(ts) \sum_{t'} \hat{P}(t'|ts) log \frac{\hat{P}(t'|ts)}{\hat{P}(t'|s)} \tag{2.10}$$

Vector Quantisation (VQ) is also used by Galata to produce a set of feature vectors corresponding to prototypical interactions. The prototypical interactions are then used to train the VLMM. The VLMM is essentially a *symbolic* predictive model, effectively meaning that the underlying continuous variables used in, say, the work of Brand on HMMs [22] are abstracted to discrete space hence becoming analogous to the finite relations in a qualitative spatial representation. This arguably produces a more understandable model since its components are higher level abstractions. The system built on this high-level semantic interpretation can be used to recognise typical interactive behaviour within the traffic domain and identify abnormal behaviour. In addition it is hypothesised that there are generative and predictive possibilities which would probably involve the learning of typical paths.

Brand's introduction of an "entropic" prior [20] which has the effect of producing a much sparser transition matrix for HMMs learned from training data has achieved some interesting results when applied to detection of human activity, in particular to a traffic intersection and people in an office. Minimisation of entropy has the effect of maximising the amount of evidence supporting each parameter while minimising uncertainty in the expected sufficient statistics between the model and the data.

There have been a number of efforts to extend and improve the standard HMM. Ghahramani and Jordan. [59] developed the factorial HMM for independent processes, Saul and Jordan [137] developed the linked HMM to model contemporaneous symmetrical processes and Jordan *et al.* [79] developed HMM decision trees. Notably Brand introduced Coupled HMMs which aim to model causally linked time series by coupling in state space, not in the outputs [19]. This has been shown to provide marked improvements over other HMM variants where there is a degree of causal interaction between datasets, for example, hand signs.

Cuzzolin *et al.* address the unsupervised case [157] again using HMMs. However the additional problem which they focus on is that of detecting actions in clutter. Given that instantiations of a particular action may not be know beforehand it is difficult to construct a model which recognises such behaviour. The models which are inferred in [157] may not necessarily be anatomically correct in the case of gait analysis, for example, but are designed to be able to identify a new instance of that action in a new scene. The essence of the approach (since this goal is not attained) is that a model of an isolated "action" can be detected in the model of a more complex action such as an action in clutter.

Related to Markov Modelling, Town [153] learns a Bayesian Network using the K2 algorithm from an "ontology". Ontology is defined by Gruber in [64]:[2]

> In the context of knowledge sharing, I use the term ontology to mean a specification of a conceptualization. That is, an ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents.

One way of interpreting this is to view the ontology as the set of descriptions allowed for a given agent i.e. the descriptive language for a particular scenario.

The ontology which Town obtains in his work comes directly from hand-labelled portions of video involving human behaviour describing the context of an individual in the video, their situations, their role and attributes. Town shows good classification results for the sequence

[2]See also `http://www-ksl.stanford.edu/kst/what-is-an-ontology.html`

from which the ontology was derived but there is no attempt to generalise behaviour beyond this.

2.3.2 Non-parametric methods

The general trend has been towards parameterised modelling of complex and large training data sets for human activity recognition, as we have seen above. There are, however, other approaches. Most notable has been the introduction of non-parametric methods where the data itself is the model. Efros *et al.* [44] showed that using the instantaneous optic flow, splitting the flow-field into four non-negative channels and blurring localised target actions could be matched between people even though they vary in size, shape and appearance. The matching in Efros' work is done by picking the nearest (in a Euclidean sense) match in the exemplar set for an newly computed set of "motion channels". The method has been demonstrated primarily on sports footage where the background is almost entirely free from clutter.

In the case of Stauffer and Grimson's work on visual surveillance [148] the claim is that over ten million objects have been tracked. From this data the aim is to:

1. obtain statistical descriptions of "normal" activity patterns,
2. detect unusual events,
3. detect unusual interactions between objects.

The interpretation of the motion tracks is performed using a classification based on the Vector Quantisation (VQ) method. VQ maps k-dimensional vectors into a finite *set of vectors*. Each of these new vectors is called a *codeword* and the set of codewords from a given set of training vectors is the *codebook*. A codeword resides in its own nearest-neighbour region and so a codeword can be chosen to represent an input vector according to the region in which the vector is placed. These codewords are referred to as *prototypes* in [148].

There are a number of algorithms for the development of codebooks of prototypes (see e.g. [58]). Stauffer and Grimson use a popular algorithm which is initialised by selecting a number of

prototypes and setting them randomly. The Euclidean distance is used as a measure to cluster the input vectors around the prototypes. A new set of protoypes is calculated by obtaining the average of each cluster. The last two steps are repeated until the prototypes do not change or the change is negligible. Each observation in a sequence is treated independently and so the probability of a given class is the product of the probabilities of that class producing each observation. Prior to classification, a codebook has been generated by the method described. The codebook can then be used as a lookup table to label new values according to the labels of nearby prototypes.

A co-occurrence matrix is then calculated which operates in a different fashion to the HMM (which detects patterns in sequences) by taking one or more observations of an object and classifying each observation into a set of classes by converting the input, be it a silhouette, an image etc., to codebook labels so that the similar objects can be placed into the same class according to the nearest prototype. The result is that the system can classify without seeing an entire sequence. Probability distributions are created from which classification of sequences can take place. Since each observation is treated as independent, the probability of a particular class is the product of that class producing each of the observations in the sequence.

While admitting that the scenes used to test the system are chosen to be well-suited to the task, the work represents a novel, probabilistic method for tracking but the classification is not as successful as the tracker. It should be noted that it is believed the classification would achieve better results by learning context cycles such as traffic light cycles which is an interesting potential use of human prior knowledge in guiding the training data to be used.

Zhong *et al.* [169] demonstrated detecting unusual activity by classifying motion and colour histograms into prototypes and using the distance from the clusters as a measure of novelty.

Non-parametric techniques for recognising the action centred on a person have become more prominent recently. In particular, the methods of Efros *et al.* [44] (which we discuss in more detail in chapter 4) and that of Blank *et al.* [14] both attempt to reliably detect actions such as walking and running regardless of where that action takes place. In fact some form of background suppression is required in both of these methods. Blank *et al.* view actions as

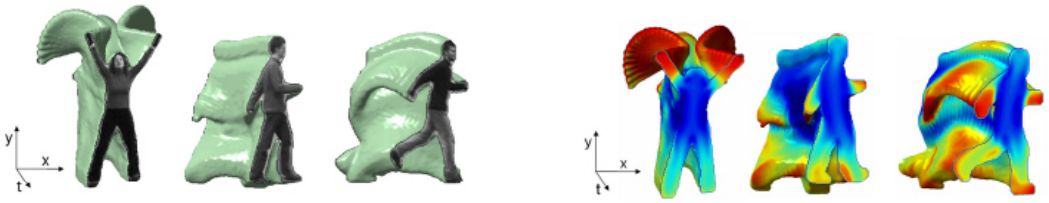


Figure 2.11: The method of Blank *et al.* represents person-centred activity as space-time shapes and uses non-parametric matching of features such as the local saliency (shown *right* with regions of high saliency encoded in red, low saliency in blue). These are computed and searched non-parametrically. Pictures are from [14].

"space-time shapes", as shown in Figure 2.11 and devise a set of features based on the solution to the Poisson equation on space-time shapes. These include space-time saliency, orientations (which are local features) and weighted moments (global features). The exemplar features vectors are then searched for the nearest-neighbour for every input action. Classification rates are impressive and the error-rate is quoted as 6.38%. The test data on which these results are generated are from sanitised video sequences and, while there is some discussion of the fact that silhouettes are not perfect, it remains to be seen how this method performs in very low-resolution situations or where the background is significantly cluttered.

2.4 Visual surveillance

Any consideration of visual surveillance could encompass an enormous variety of aspects including sensors, photogrammetry etc. However it is approached here in the context of learning, detection and explanation of human-initiated behaviours, which is the area of most relevance to this work.

It has been said by Picard:

> Successful biases must happen at both the low and high levels, and depend on both the data and the goals [117].

The preceding review of action and behaviour recognition, from an Artificial Intelligence standpoint, corresponds to "low-level" analysis in that it deals with data arising directly from the sensors. In the computer vision literature, unsurprisingly, the majority of the work falls into

this category. The "high-level" is concerned with interpretation of the observed activity and, while this is a theme more readily found in the AI literature, there is a distinct lack of examples where researchers have attempted to join the low and the high level.

In surveillance, it is typically expected that the physical area being monitored will, to a greater or lesser degree, be known. This is not to say that each and every type of behaviour which a human could classify as unusual or suspicious can be determined in advance. In fact to attempt such a task could be considered an intractable problem unless the environment in question has a highly constrained rule set governing the actions of the objects in the scene.

In automatic surveillance the goal is to determine which type of action is being observed at a given time and make an informed judgement about the the behaviour. *Is it a threat? What is happening?* However, human behaviour is complex and the potential interactions between humans in even the most common real-life environment, such as an urban street, increases the complexity dramatically. Learning all possible examples of interactions in these scenes is not feasible, especially using current techniques such as HMMs which require many examples of each interaction. Therefore we propose to approach the problem not as one of classification by exhaustively large datasets but as one of classification with the aid of important prior knowledge about the scene.

There has been much reported in the recent literature about methods for training recognition systems using large training data sets (see, for example, [158]). This approach is beneficial in the case where not much is known about the class of object that one wishes to detect. In this work we know we want to detect behaviour which deviates from "normal" and classify behaviour which does not. Humans can do this very effectively; for example, Imagery Analysts in the military domain have no trouble detecting troop or vehicle formations from very sparse doppler-shift radar returns mainly due to the enormous range of prior knowledge at their disposal.

Nairac and Tarassenko are proponents of the idea that learning normality alone is all that is required for the detection of abnormality [103, 152, 102]. Their work has shown that, using a number of different similarity measures, it is possible to reliably detect unusual behaviour, for example behaviour which can lead to crucial failure of an aircraft engine, using neural

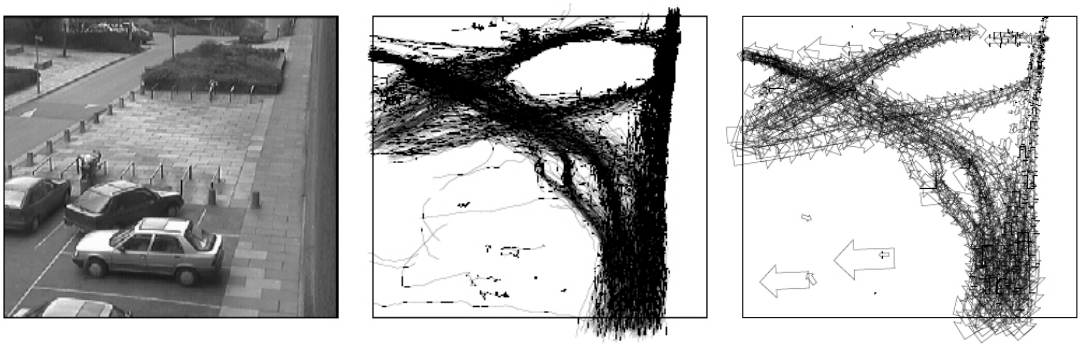


Figure 2.12: The work of Johnson and Hogg on learning patterns of motion in a pedestrian scene (*left*), a large amount of trajectory training data is used (*centre*) to compute a number of prototypical motions (*right*). These figures are from [78].

networks. This method of novelty detection on the basis of a representation of normality learnt exclusively from normal examples has been applied to a number of areas ranging from medical image analysis to engineering component monitoring.

The work of Johnson and Hogg in [78] is a clear attempt to introduce the concept of action and behaviour into classification systems resulting from computer vision methods. Moreover the idea that object shape is significant is addressed in Johnson and Hogg's work but shape and trajectory information is not maintained independently.

An example in which human-level descriptions of video have been derived is that of Gerber, Nagel and Schreiber [56] where a traffic scene is analysed and traffic queues reported using textual descriptions (see Figure 2.14. (In fact, traffic scene analysis is a predominant theme in high-level behaviour analysis see e.g. *Towards Robust Automatic Traffic Scene Analysis* [85].) The work of Gerber *et al.* is based on a vision subsystem to provide tracks of cars in the scene, a model of the traffic lanes which is generated by hand and a rule-based system (using Fuzzy Metric-Temporal Horn Logic) which searches for an interpretation of the "facts" observed using the cameras. The resulting output gives information regarding how many cars were in the queue, when the queue formed and so on.

Similar attempts for scenes with more direct human activity, i.e. where the person can be seen, have been less prominent. Mann *et al.* [93] did, however, report on progress towards recognising and explaining interactions between people (dynamic) and objects (static). This is an extension of the work by Ikeuchi and Suehiro [73] and Siskind [147] to consider both

Figure 2.13: The traffic analysis of Gerber *et al.* produces text descriptions of traffic queues and uses high-level markup of the scene. On the left is the objects being tracked. On the right the traffic lanes have been identified by hand. These pictures are from [56].

kinematic and dynamic properties in time-varying scenes containing rigid objects. Given a fairly rich descriptive language the authors demonstrate a number of feasible interpretations of the activity "lifting a Coke can" can be achieved and these are ordered in preference showing there is knowledge of uncertainty in the system.

Kingston University have expended considerable effort in creating solutions for a deployable, wide-area visual surveillance system and have addressed a variety of issues including themes such as colour constancy [122] and learning semantic models [91] in addition to the more common problems associated with surveillance e.g. tracking. Notable work includes the investigation of how to track through blind regions i.e. areas between camera views which cannot be seen in any view [12]. Statistics for camera handover with a Kalman Filter tracking in 3D are 87.3% overall but for blind regions with a temporal gap of 2 seconds the success rate drops to 30%. The method developed by Black *et al.* in [12] is based on an agent-based tracker which manages multiple hypotheses for entry and exit points between cameras and the authors report a rate of 90% success in the handover of tracks between multiple cameras with significant blind regions [51].

Xiang and Gong have addressed an important issue which is central to this thesis: how to effectively recognise action in a surveillance context when there is a sparsity of example data [166] and what rôle the high-level labelling of trajectories plays in this situation and in the general case [167]. The motivation given by Xiang and Gong for the preference for unlabelled classification is: (1) "Manual labelling is laborious", (2) "Manual labelling ... could be inconsistent or error prone", and (3) "... training using labelled data does not necessarily help a

model with identifying novel instances of atypical behaviour patterns . . . ". In [166] the authors consider a particularly limited set of data and it is not clear from these preliminary results that the labelling of tracks results in manifestly worse activity classification (when the classes are simply "normal" or "abnormal") *because* the novel classification method reported has greater "insight" into the notion of normality and abnormality. The main drawback of Xiang and Gong's work is that the authors do not address the most obvious criticism which is that the labelling the authors use for the training data could be innacurate. Referring to (2) above, this is an expectation the authors seem to have and, therefore, new examples do not classify well due to the subtle variations in position and speed which are observed. This latter point strongly suggests that further work is required to clarify the interesting question of whether manual labelling is an aid to human-activity classification, especially since trajectory data *alone* was used as the feature vector. This criticism is further corroborated by the fact that, in the only reported work addressing explicitly the issue of manual labelling, a single, very simple, scenario was used in the analysis.

From an Artificial Intelligence standpoint the AI Lab at MIT has developed an entirely automated system for visual surveillance and monitoring of an urban site. This work, mentioned above, [63] incorporates automatic tracking using an adaptive background model and classifies activity and objects using a co-occurrence measure based on the trajectories. While this is an interesting example of engineering computer vision solutions it does not attempt to *explain* behaviour. More recently there have been moves towards explanation of behaviour in video with attempts to describe and query video at the action and not the feature level [84]. This work involves combining research on Question Answering and Natural Language Understanding [83] with Computer Vision. The system presented uses surveillance video as an example and can answer questions at the level of, "Did any cars leave the garage?".

Highly complex scenarios, specifically interaction in crowds, are currently being investigated. Notably the global behaviour of the crowd, as opposed to the individuals that comprise the crowd, is the focus of this work [1, 2] i.e. optic-flow of the group is used rather than single-person tracks. The immaturity of this area is demonstrated by the fact that the initial research is

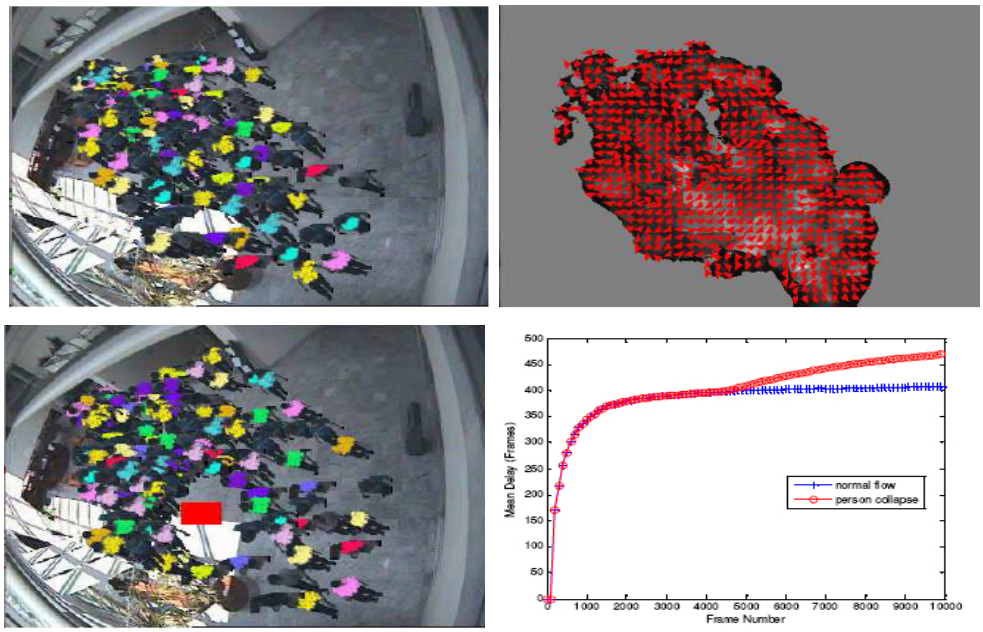
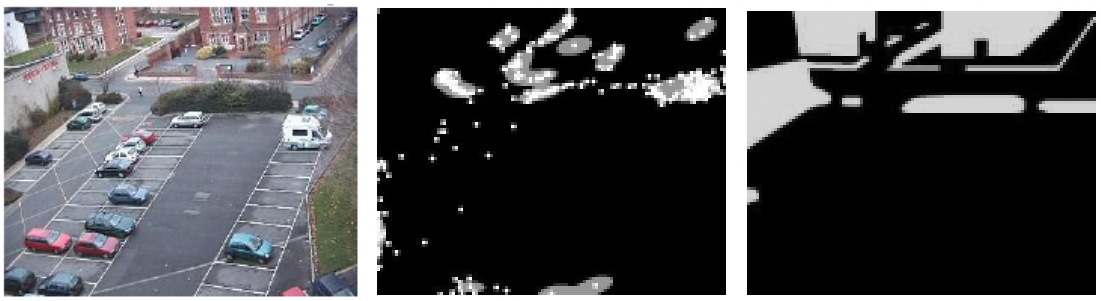


Figure 2.14: Andrade and Fisher use synthetic data to generate anomalies among crowds of people. *Top left*: the person agents are programmed to adhere to a certain trajectory and avoid other agents; *Top right*: The optic flow field is computed for the normal crowd behaviour shown in the *top left* figure; *Bottom left*: A blockage is introduced and the crowd behaviour is generated such that the agents avoid the obstacle and crush one another; *Bottom right*: The mean time taken for each agent to leave the scene is observed to increase for the scenario where a blockage is introduced. Figures are from [1].

focused on modelling crowd behaviour using a person software agent model of social interaction. Andrade and Fisher [1, 2] generate hypothetical anomalies within the 'crowd of agents' (such as when a person falls to the ground or when a crush is in evidence at e.g. a barrier) which could then be used (it is suggested) for training purposes or validation of a computer system.

The work of Dee and Hogg [39] is particularly interesting to the aims of this thesis. They develop a novel method for detecting "inexplicable" behaviour which is based upon a model of how humans navigate a scene towards a "goal". See Figure 2.15 for examples of potential goals within an urban scenario. There is no statistical representation of normal activity in Dee and Hogg's work. The model of human activity takes as its input the headings of agents within the scene and a labelled scene map. From the headings and the obstacles obtained from the scene map, area *visibility* is derived. A Markov chain is predefined which enables a score to be computed which penalises certain transitions which an agent can make. The goal with the lowest predicted score is taken as the most likely explanation for the behaviour of the agent at any time step. These scores are compared to how unusual a human finds the activity and a



(a) (b) (c)

Figure 2.15: The algorithm of Dee and Hogg for detecting unusual activity relies on high-level markup of the scene (from [39]). (a) The view of the scene from a static camera, (b) the exits, (c) the obstacles.

correlation between the human interpretation and that of the system is demonstrated.

As far as human-intervention in this goal-directed system is concerned, Dee and Hogg demonstrate that scene exits can be automatically recovered directly from the statistics of the tracks over a long period of time. The obstacles, however, are hand-defined, however, as must stationary objects such as parked cars, as shown in Figure 2.15.

One weakness of this work is the assumption that general body-direction is the prime indicator of visibility. This is particularly significant given that the model is attempting to achieve *goal*-directed analysis. That is, explicitly, *intention* is being modelled. In many situations, including those examples shown in Dee and Hogg's work, this assumption is fair. We can easily think of instances where gaze-direction is a better cue to intention than body-direction alone, however.

In summary this work is a very interesting example of successfully exploiting human prior knowledge about how people *ought to* behave in specific urban locations. It is not clear, though, to what extent this approach is scalable, due to the simplistic assumptions required to infer where the goal is located. That is, there may be reasons other than obstacles which prevent a person moving directly towards a goal. For example, another agent in the scene may be behaving in a dangerous manner and should therefore be avoided. In such an example, knowledge of other agents and higher-level goal representation would be required.

2.5 Conclusion

The dominant theme in computer vision in the technical approaches to the recognition of human activity in video sequences is to learn parameterised models from training datasets representing normality. The training data is often nothing more complex than trajectories i.e. position and velocity. The modelling methods are typically Bayesian with Hidden Markov Models being a recurring theme in the published literature. Sophisticated techniques have been implemented to solve the problem of incorporating 3D views from multiple cameras, for example, and complex statistical techniques for tracking.

Although there is evidence of a degree of supervision in the learning processes in the literature, where it occurs it tends to take the form of supervising *what* is learned as opposed to *how* it is learned. Humans use a combination of prior knowledge and learning when analysing visual information but there is little evidence, beyond the application of Bayes' Rule which clearly accounts for evidence and prior beliefs, of a serious attempt to develop surveillance algorithms that can fully utilise expert knowledge.

The AI community has focussed on this issue but, as Rigolli and Brady point out [124], the development of the ontology for reasoning about the human-motivated activity is often divorced from the information which can actually be obtained from the sensors.

And so there exists a gap which requires to be filled if "intelligent surveillance" - the analysis of dynamic human activity in real-world scenarios which achieves human-level recognition and can reason about the activity - is to become a reality. We propose on the basis of the scientific state-of-the-art that the following issues need to be considered:

1. How to incorporate high-level expert knowledge in a vision-based system.
2. When training data is sparse, learning methods become increasingly unreliable. In many scenarios, especially in the military domain, there is a lack of exemplars of action types. What methods are appropriate in this case?
3. Looking beyond trajectories, what other information can be extracted from surveillance

video which would expand the capabilities of a system for recognising human behaviour?

3

Gaze estimation in video

In this chapter we describe a new method for estimating where a person is looking in images where the head of a person is low-resolution, typically 20 pixels high. The lowest-level of our method is a feature vector which is based on skin detection. This feature is used to estimate the pose of the head, which is discretised into 8 orientations relative to the camera. A fast sampling method returns a distribution over head pose relative to a camera centred frame. The overall body pose relative to the camera frame is approximated using the velocity of the body, obtained via colour-based tracking in the image sequence. We show that, by combining direction and head pose information, gaze direction is determined more robustly than using each feature alone. We demonstrate this technique on surveillance and sports footage.

The results of this chapter have been published in the Proceedings of The European Computer Vision Conference (ECCV), Graz, Austria, 2006 [132] and "Human Activity Recognition and Modelling" at the British Machine Vision Conference (BMVC), Oxford, 2005 [130].

3.1 Introduction

In applications where human activity is under observation, be that CCTV surveillance or sports footage, for example, knowledge about where a person is looking (i.e. their gaze) provides observers with important clues which enable accurate explanation of the scene activity. It is possible, for example, for a human readily to distinguish between two people walking side-by-side but who are not "together" and those who are acting as a pair. Such a distinction is possible when there is regular eye-contact or head-turning in the direction of the other person. In soccer or rugby, for example, head position is often a guide to where the ball will be passed next i.e. it is an indicator of *intention*[1] which is essential for causal reasoning. In this chapter we present a new method for automatically inferring gaze direction in images where any one person represents only a small proportion of the frame, where the head ranges from 20 to 40 pixels high.

The first component of our system is a descriptor based on skin colour. This descriptor is extracted for each head in a large training database and labelled with one of 8 distinct head poses. This labelled database can be queried to find either a nearest-neighbour match for a previously unseen descriptor or, as we discuss later, is non-parametrically sampled to provide an approximation to a distribution over possible head poses.

Recognising that general body direction plays an important rôle in determining where a person can look due to anatomical limitations, we combine direction and head pose using Bayes' rule to obtain the joint distribution over head pose and direction, resulting in 64 possible gazes, since head pose and direction are discretised into 8 sectors each as shown in Figure 3.1.

The chapter is organised as follows. First, we highlight relevant work in this, and associated, area(s). We then describe how head-pose is estimated in section 3.3. In section 3.4 we provide motivation for a Bayesian fusion method by showing intermediate results where the best head-pose match is chosen and, by contrast, where direction alone is used. Section 3.4 also discusses how we fuse the relevant information we have at our disposal robustly to compute a

[1] In all but the most highly-skilled teams, where awareness of a team-mate's position appears to be intuitive, a player at least glances in the direction of the intended pass.

distribution over possible gazes, rejecting non-physical gazes and reliably detecting potentially significant interactions. Throughout the chapter we test and evaluate on a number of datasets and additionally summarise comprehensive results in section 3.5. We conclude and discuss potential future work in section 3.6.

3.2 Review of relevant literature

Determining the instantaneous focus of a person's attention in surveillance images is a challenging problem that seems to have received no attention until now. In fact this problem was first addressed by us very recently [130, 132].

Everingham and Zisserman [46] did, interestingly, develop a technique for overlaying 3D head models on faces, with a resolution in the range 15 to 200 pixels high as a means to identifying people in broadcast video sequences. This could have potentially been extended to determine where the person is looking but the crucial drawback with Everingham and Zisserman's work in relation to surveillance video is the fact that they search for faces of a *specific* character whose appearance is known *a priori* and for whom a 3D face model has been constructed in advance. This would clearly be impossible in a surveillance application where nothing is known about the appearance of the person under observation before they appear in the video.

Closely related in technical approach to the work of this chapter is that of Efros *et al.* [44] for recognition of human action at a distance. That work showed how to distinguish between human activities such as walking or running by comparing gross properties of motion using a descriptor derived from frame-to-frame optic-flow and performing an exhaustive search over extensive exemplar data. Head pose is not discussed in [44] but the use of a descriptor invariant to lighting and clothing is of direct relevance to head pose estimation and has inspired aspects of our algorithm.

Dee and Hogg [39] developed a system for detecting unusual activity which involves inferring which regions of the scene are visible to an agent within the scene. A Markov Chain with penalties associated with non-hidden state transitions is used to return a score for observed

trajectories. The state transition penalties essentially encode how directly a person made his/her way towards predefined goals, typically scene exits. In their work, gaze inference is vital, but gaze is inferred from trajectory information alone which can lead to significant interactions being overlooked, as we show later in this chapter, because the assumption that the head is always aligned with body-direction is not robust.

In contrast, there has been considerable effort to extract gaze direction from relatively high-resolution faces, motivated by the drive toward ever better Human/Computer Interfaces. The technical aspects of this work have often focused on detecting the eyeball primarily. Matsumoto and Zelinsky [94] compute 3-D head pose from 2-D features and stereo tracking. Perez et al. [115] focus exclusively on the tracking of the eye and determination of its observed radius and orientation for gaze recognition. Kaminski et al. [80] have achieved a very similar goal but using a single image while retaining a face and eye model. Gee and Cipolla's [54] gaze determination method based on the 3D geometric relationship between facial features was applied to paintings to determine where the subject is looking. Related work has tackled expression recognition using information measures. Shinohara and Otsu demonstrated that Fisher Weights can be used to recognise "smiling" in images. Osadchy et al. use a neural network to detect faces and estimate pose similtaneously [106]. It must be noted that, even for the higher-resolution images expected in an HCI-type application, non-frontal face detection is still unreliable.

While this approach is most useful in HCI where the head dominates the image and the eye orientation is the only cue to intention, it is too fine-grained for surveillance video where we must usually be content to assume that the gaze direction is aligned with the head-pose i.e. one cannot track the eyes. In typical images of interest in our application area (low/medium resolution), locating significant features such as the eyes, irises, corners of the mouth, etc as used in much of the work above is often infeasible. Furthermore, though standard head/face-detection techniques [159] work well in medium reolution images, they are much less reliable for detecting, say, the back of a head, which still conveys significant gaze information. Methods such as "AdaBoost" have been applied to non-frontal face-detection i.e. profiles [100, 139] but the 360^o head-pose detection problem has not been solved with these techniques. Indeed

performance is not robust for face-detection when the pose of the face can vary significantly. The "Eigenface" [154] or "Fisherface" [9] methods require that the input images are registered with fairly high precision which is impossible to achieve across pose variations. View-based approaches have taken the approach of representing the face using a separate model for each of a limited set of poses [114, 34]. 3D model approaches have used standardised face databases [135, 15] and there is little reported work on less constrained views such as those found in TV or surveillance footage.

The lowest level of our approach is based on skin detection. Because of significant interest in detecting and tracking people in images and video, skin detection has naturally received much attention in the Computer Vision community [27, 72, 76]. However skin detection alone is error-prone when the skin region is very small as a proportion of the image. That said, contextual cues such as direction can help to disambiguate gaze using even a very coarse head-pose estimation. By combining this information in a principled i.e. probabilistic, Bayesian fashion, gaze estimation at a distance becomes a distinct possibility as we demonstrate later in the chapter.

The aim is to determine which pixels in an image correspond to skin and non-skin. Perhaps the most straightforward method is to construct a look-up table by deciding in advance in which regions of a given colour-space skin colour is found. This method was used by Chai and Ngan [27]. This technique is unreliable in medium-scale images, however. Hidai et al. [72] defined an ideal skin colour using an average of exemplar face images from which they defined skin and non-skin pixels via non-parametric matching. Parameterised techniques usually involve multi-variate Gaussians, the parameters of which are learned using the Expectation-Maximisation algorithm (see e.g. [76]).

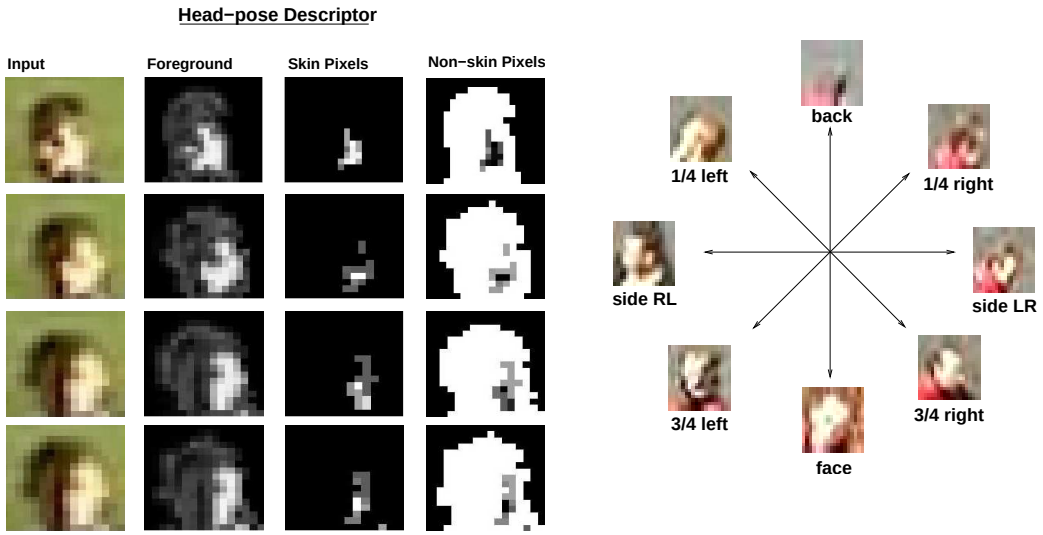


Figure 3.1: The figure on the left shows the images which result from the mean-shift image patch tracker (*col. 1*) (with an additional step to stabilise the descriptor by centring the head in the window), subsequent background subtraction (*col. 2*), the weight image which represents the probability that each pixel in the head is skin (*col. 3*) and non-skin (*col. 4*) (non-skin is significant as it captures proportion without the need for scaling). The concatenation of skin and non-skin weight vectors is our feature vector which we use to determine eight distinct head poses which are shown and labelled on the right. Varying lighting conditions are accounted for by representing the same head-pose under light from different directions in the training set. The same points on the "compass" are used as our discretisation of direction i.e. N, NE, E, etc.

3.3 Head pose detection

3.3.1 Head pose feature vector

Although people differ in colour and length of hair and some people may be wearing hats, beards etc. it is reasonable to assume that the amount of skin that can be seen, the position of the skin pixels within the frame and the proportion of skin to non-skin pixels is a relatively invariant, if coarse, cue for a person's gaze direction in a static image. We obtain this descriptor in a robust and automatic fashion as follows. First, a mean-shift tracker [31] is automatically initialised on the head by using naive background subtraction to locate people and subsequently modelling the person as distinct "blocks", the head and torso. Second, we centre the head within the tracker window at each time step which stabilises the descriptor ensuring consistent position within the frame for similar descriptors. That is, the head images are scaled to the same size and, since the mean-shift tracker tracks in scale-space we have a stable, invariant, descriptor.

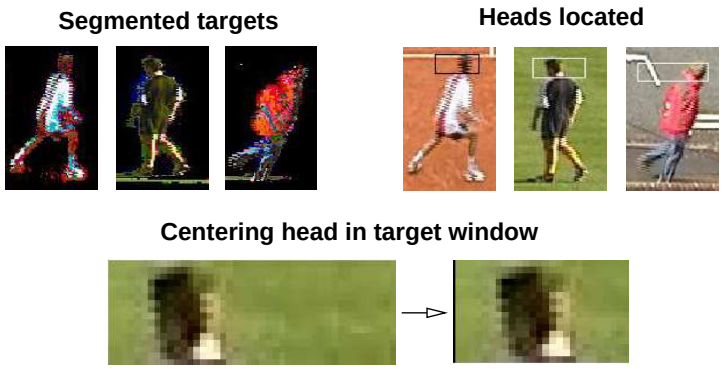


Figure 3.2: Automatic location of the head is achieved by segmenting the target using background subtraction (*top-left*) and morphological operations with a kernel biased towards the scale of the target to identify objects. The head is taken as the top 1/7th of the entire body (*top-right*). The head is automatically centred in the bounding box at each time step to stabilise the tracking and provide an invariant descriptor for head pose, as shown in the second row.

Third, there is no specific region of colour-space which represents skin across all sequences and therefore it is necessary to define a skin histogram for each scenario by hand-selecting a region of one frame in the current sequence to compute a normalised skin-colour histogram in RGB-space. (It has been demonstrated that there is no difference in the performance of skin detectors based on colour-regions when RGB or YCbCr, HSV etc. colour-spaces are used [116].) We then compute the weights for every pixel in the stabilised head images which the tracker automatically produces to indicate how likely it is that it was drawn from this predefined skin histogram[2]. Using the knowledge of the background we segment the foreground out of the tracked images. Every pixel in the segmented head image is drawn from a specific RGB bin and so is assigned the relevant weight which can be interpreted as a probability that the pixel is drawn from the skin model histograms.

Some meanshift implementations suggest a histogram discretised into 20 bins for each dimension of colour space. So if a 3-D histogram is computed with axes along the R, G and B dimensions of the colour-space then the histogram is an 8000-element volume. The actual skin-colour occupies a very small region of this volume. A significant amount of computational effort is therefore expended computing this large histogram for each step of the tracker since the weights are computed at each frame.

[2]This will be recognised as a similar approximation to the Battacharyya coefficient as implemented in the meanshift algorithm [31].

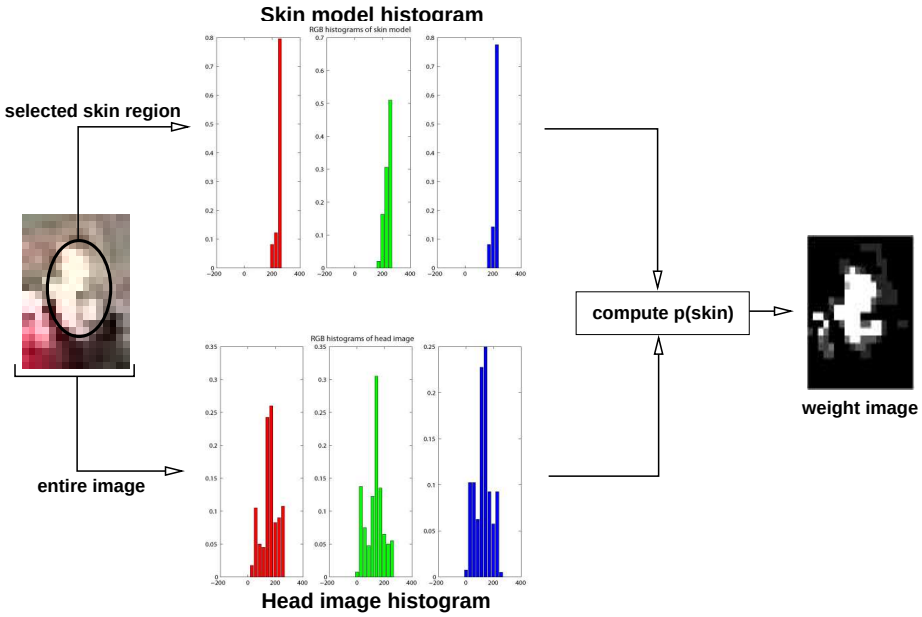


Figure 3.3: The R,G and B histograms of the skin model occupy a small amount of the colour-space. The face region from which the histograms are computed is shown as the region inside the ellipse superimposed on the face (*left*).

We split the RGB space into three independent histograms, compute the likelihood that each pixels R, G and B value was drawn from that histogram and multiply together to obtain a likelihood that each pixel was drawn from the overall (RGB) skin histogram. For every bin i (typically we use 10 bins) in the predefined, hand-selected skin-colour histograms q_R, q_G and q_B the histograms of the tracked image p_R, p_G and p_B a weight, w_i, is computed:

$$w_i = \sqrt{\frac{q_{R,i}}{p_{R,i}}} \cdot \sqrt{\frac{q_{G,i}}{p_{G,i}}} \cdot \sqrt{\frac{q_{B,i}}{p_{B,i}}} \tag{3.1}$$

Every foreground pixel in the tracked frame falls into one of the bins according to its RGB value and the normalised weight associated with that pixel is assigned to compute the overall weight image, as shown in Figure 3.1. The non-skin pixels are assigned a weight that the pixel is *not* drawn from the skin histogram. This non-skin descriptor is necessary because it encodes the "proportion" of the head which is skin, which is essential as people vary in size and scale. Each descriptor is scaled to a standard 20×20 pixel window to achieve robust comparison when the head sizes vary. Finally, in order to provide temporal context to our descriptor of head-pose we concatenate individual descriptors from 5 consecutive frames of tracker data for a particular example and this defines our instantaneous descriptor of head-pose.

3.3.2 Training data

Algorithm 1 To obtain head-pose training data

1: Track head in a video sequence
2: Centre head within tracker window at each frame
3: Define skin histogram for sequence (by hand, if necessary)
4: Segment the foreground in every image
5: For every pixel belonging to the foreground compute p(skin) and p(non-skin)
6: Concatenate 5 frames of each feature vector per frame

We assume that we can distinguish head pose to a resolution of 45^o. It is unlikely that the coarse target images would be amenable to detecting head orientations at a higher degree of accuracy in any case. This means discretising the 360^o orientation-space into 8 distinct views as shown in Figure 3.1. The training data we select is from a surveillance-style camera position and around 100 examples of each view are selected from across a number of different sequences and

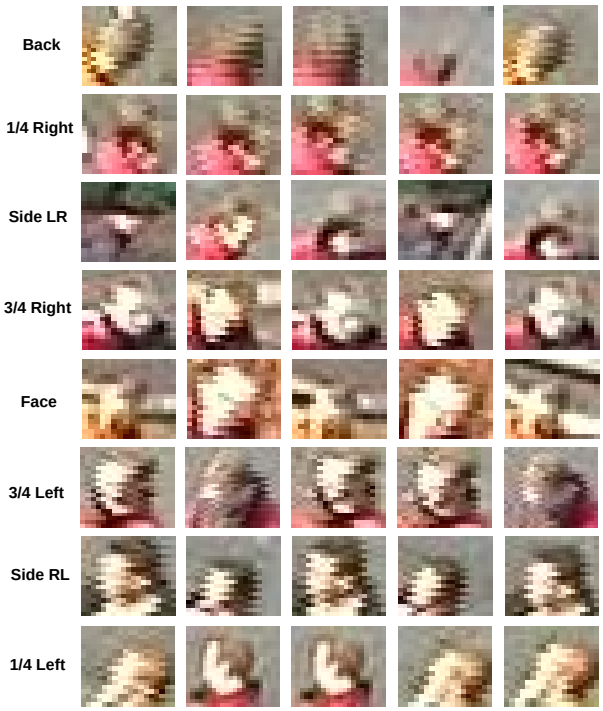


Figure 3.4: Example training data from each of the 8 discretised views shown in Figure 3.1. Note that the data, both for the training database and the input sequences used to test the method, is used as it exists, with no pre-processing. So, in particular, as seen in some examples here, interlacing effects may be apparent. In practise, this poses no evident difficulties for our algorithm.

under different lighting conditions i. e. light from left, right and above. This is because, as can be seen in Figure 3.1 there is a tendency for the skin pixels to be identified to track saturated pixels, whether by the user in the histogram selection stage or at the pixel weight computation. It was not found that using an intensity-invariant colour-space solved this problem. The head was automatically tracked as described above and the example sequence labelled accordingly. The weight image for 5 consecutive frames are then computed and this feature vector stored in our exemplar set. The same example set is used in all the experiments reported e.g. there are no footballers in the training dataset used to compute the gaze estimates presented in Figure 3.18. It is necessary, therefore, that the training data be from a comprehensive set of lighting conditions due to varying directions and strength of illumination.

3.3.3 Matching head poses

The descriptors for each head pose are $(20 \times 20 \times 5 =)2000$ element vectors. With 8 possible orientations and 100 examples of each orientation searching this dataset rapidly becomes an

issue.

Beis and Lowe reported a variant on the k-d tree algorithm [8] and Nene *et al* [105] proposed an algorithm which uses a Euclidean distance measure but is efficient for dimensions greater than 15, where most algorithms are impractical. Matching has also been performed using wavelet coefficients [4, 75] and various pyramid representations [38, 66, 164].

McNames provided an overview of a number of common algorithms' performance which demonstrated that a Principal Components Tree search outperforms the other well-known methods [97].

Therefore, we elect to structure the database using a binary-tree in which each node in the tree divides the set of exemplars below the node into roughly equal halves. Such a structure can be searched in roughly $\log n$ time to give an approximate nearest-neighbour result. We do this for two reasons: first, even for a modest database of 800 examples such as ours it *is* faster by a factor of 10; second, we wish to frame the problem of gaze detection in a probabilistic way and Sidenbladh [144] showed how to formulate a binary tree search in a pseudo-probabilistic manner. This search is based on the sign of the Principal Components of the data, as we illustrate in Figure 3.5. This technique was later applied to probabilistic analysis of human activity in [131]. We provide some detail on this method in the following section.

3.3.4 Database creation and search

In [144] a large database of high-dimensional points is structured as a binary tree via principal component analysis of the data set. The children of each node at level i in the tree are divided into two sets: those whose i^{th} component (relative to the PCA basis) is larger and those whose value is smaller than the mean. In Sidenbladh's application each data point comprised the concatenated joint angles over several frames of human motion capture data. The method, however, applies equally well to our application of image feature data and the pseudo-random search algorithm is identical to that derived in [144].

If $\bar{\Psi}$ is a length dm vector representing the median of all the sequences of head-pose descriptors

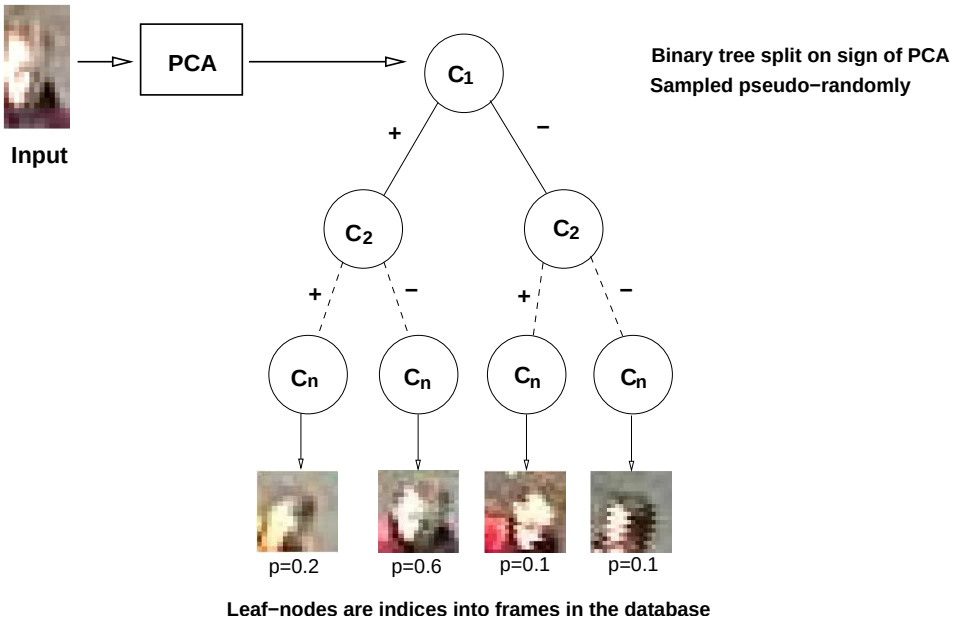


Figure 3.5: This figure illustrates how we sample from the databases to produce a distribution over the exemplar data given some input descriptor. In this case, the descriptor is the probability of skin/non-skin pixels in the face image shown at the top-left. The PCA decomposition of the descriptor (concatenated over 5 frames to provide some temporal context) is used to decide how to traverse the binary tree, branching depending on the sign once the median has been subtracted (to balance the tree). At each branching of the tree a randomness factor is computed (based on a Gaussian) which results in the leaf nodes of the tree being explored. The leaf nodes are indices into the database which, in turn, point to specific frames in a sequence. We show here illustrative the matches generated for 10 samples with associated probabilities.

(the skin/non-skin feature vectors) i.e.

$$\bar{\Psi} = \frac{1}{n}\sum_{i=1}^{n}\Psi_i \tag{3.2}$$

and

$$\hat{A} = [\hat{\Psi}_i, \ldots, \hat{\Psi}_n] \tag{3.3}$$

is a $dm \times n$ matrix containing all the sequences with the median of the entire set of training

descriptors subtracted, by applying Singular Value Decomposition we write

$$\hat{A} = U\Sigma V^T \tag{3.4}$$

where the $dm \times n$ matrix U contains the principal components of $\hat{A}$ and Σ is diagonal matrix containing the standard deviation σ_l accounted for by the principal components $l = 1, \ldots, n$. Any sequence in the database can be approximated by

$$\Psi_{match} = \bar{\Psi} + U\mathbf{c}_{match} \tag{3.5}$$

Where $\mathbf{c}_{match}$ is the sampled nearest-neighbour match from one traversal of the binary tree.

Significantly, the first $b = log_2(n)$ (where n is the number of time intervals in the training data) components are selected.

(If $n \approx 50000$ and $b = 16$ this accounts for 89% of the variance in the training data i.e.

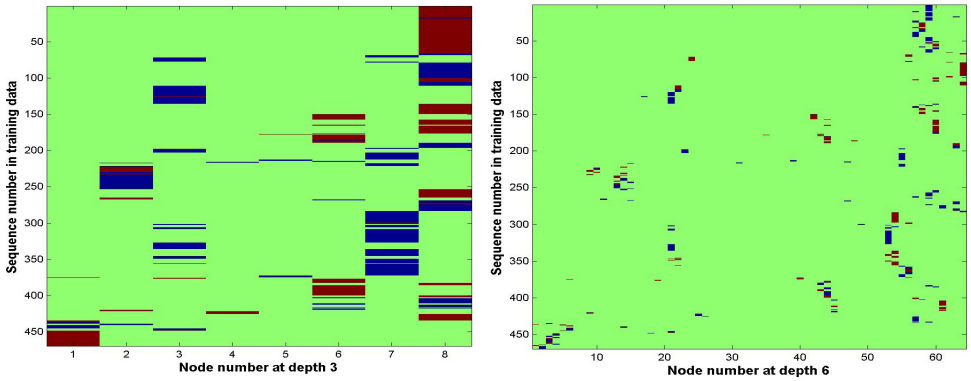
$$\frac{\sum_{l=1}^{b} \sigma_l^2}{\sum_{l=1}^{n} \sigma_l^2} \geq 0.89.) \tag{3.6}$$

These components are then organised into a binary tree the nodes of which are split on the basis of the sign of the components once the median value has been subtracted:

$$\mathbf{c_i} = [c_{i,1}, \ldots, c_{i,b}] \tag{3.7}$$

The search of the tree is randomised by the inclusion of a random perturbation of the traversal of the tree drawn from a Gaussian distribution. That is, it is decided which branch of the tree to choose, at each level l for the Principal Component coefficient at that node $c_{t,l}$ and the input coefficients at that level, $c_{i,l}$, based on the probabilities:

$$p_{right} = p(c_{t,l} \geq 0 | c_{i,l}) = \frac{1}{\sqrt{2\pi}\sigma_l} \int_{z=-\infty}^{c_{t,l}} \exp^{-\frac{z^2}{2\sigma_l^2}} dz \tag{3.8}$$



(a) (b)

Figure 3.6: We show here how the database has been constructed for the person-centred action (blurry optical flow) feature set. Each image represents one level of the binary tree with the indices into frames on the y-axis and the node on the x-axis. The shaded blocks represent the occupancy of that frame at that node, depending on how the tree has been split (blue = 0, red = 1). Green represents zero meaning no exemplar coefficients reside at that node. The nodes shown are (a) depth 3 and (b) depth 6. This demonstrates the tree is fairly evenly split, which is important for traversal when searching.

$$p_{left} = 1 - p_{right} \tag{3.9}$$

At the leaf nodes a linear search takes place if there is more than one match. The probability of these matches is computed on the basis of how "close" the match in the database is to the input i.e.

$$p(match|input) \propto \exp -(\frac{|match - input|}{\sigma})^2 \tag{3.10}$$

This search method is used for two reasons: it is more efficient and the ability to return multiple neighbours represents a distribution over possible actions i.e. a likelihood. The search time is improved by a factor of 20 and, since we sample many times, the search provides a set of particles which represents a distribution over the exemplar feature vectors into frames of the previously seen examples.

We achieve recognition rates of 80% (the correct example is chosen as the ML model 8/10 queries) using this pseudo-probabilistic method based on Principal Components with 10 samples. For comparison, we show the statistics for a linear search of the complete feature set and for a linear search on the PCA components derived from the features in the table in Figure 3.9.

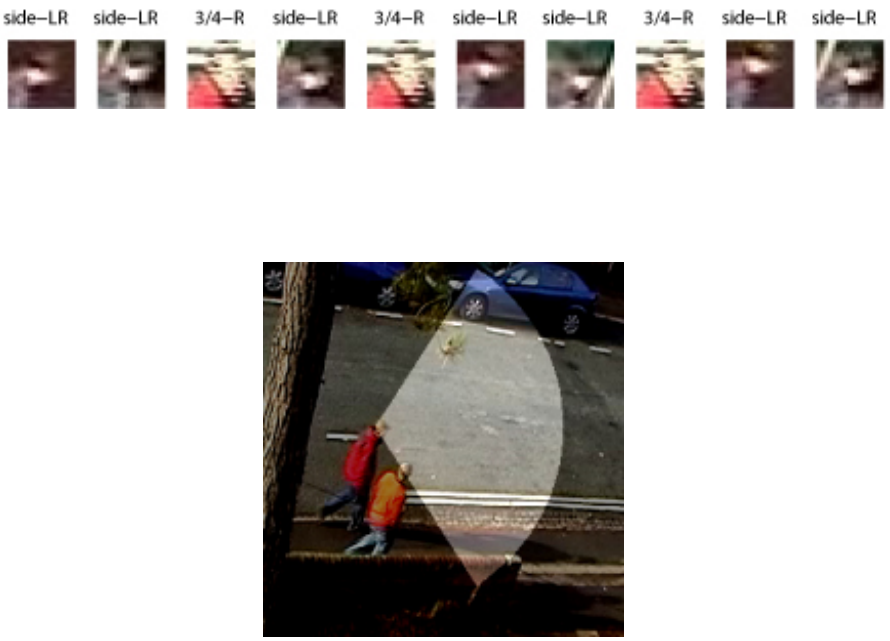


Figure 3.7: (*Top*) the best-matching frames for each of 10 samples of the database are shown here. From the head-pose labels a distribution over best-matching head-poses is computed and the ML head-pose is selected, which is shown here (*bottom*) superimposed on the original frame.

Figure 3.8: The distribution over head-poses resulting from the 10 best results for a linear search of the database for this input frame is shown here. The linear search, even on the PCA coefficients is slightly more robust in terms of finding the best matching example (see Figure 3.9). But for databases of the scale of that used here it is considerably slower. We show in Figure 3.9 that it is 20× slower for 10 samples of our head-pose feature database.

Search type	**Detection rate (%)**	**Search time per sample (secs)**
Nearest-neighbour (full data)	83.2	0.461
Nearest-neighbour (PC coeffs)	81.9	0.426
Sampling (per sample)	77.9	0.023

Figure 3.9: Comparison of detection rate for three types of head-pose matching search. As expected full comparison of the input descriptor (first row) gives best results with comparison using the Principal Components giving similar results. The sampling method described in the text returns a distribution over possible matches and the figures quoted are for the frequency of ML match corresponding to a true match and when a match is found in the distribution but not necessarily the ML match. While detection rate is inferior the probabilistic information can be exploited and the search is considerably faster

.

Figure 3.10: Detecting head pose in different scenes using the same exemplar set.

An illustrative example of the distribution over the database given an input head image in this is shown in Figure 3.7. Results of sampling from this database for a number of different scenes are shown in Figure 3.10.

3.3.5 Rectification to the ground-plane

Gaze inference is only of use if we can conclude from the estimation of gaze what it is that a person can see or, even better, what he/she is *looking at*. The human visual system has a field-of-view (FOV) of 105^o [119]. Picking an arbitrary visual range therefore allows the $2D$ visual field to be drawn on the images. Note that there is no occlusion reasoning in the system so this is an idealised indication of what can be seen. What can truly be seen by the person is in the world and not the image plane. Therefore we must invest some effort to correct for various perspective effects.

Computing a planar homography

The homography computation allows the image to be "ortho-rectified". That is, to warp the original image in such a way that the view is as though the image was capture by a camera whose

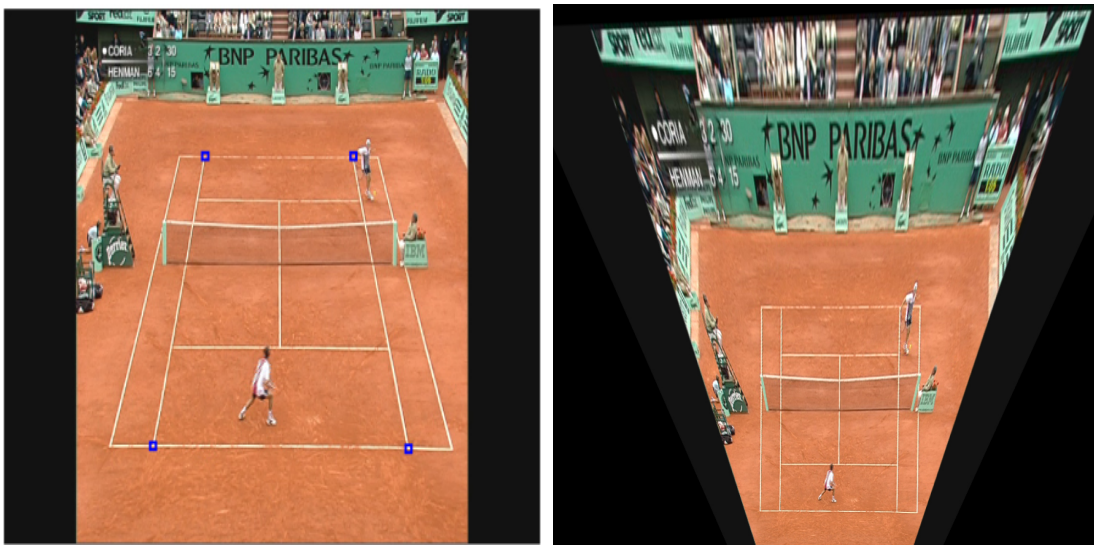


Figure 3.11: The rectification of an image to the ground plane is achieved by computing the projective transform between point correspondences. The control points shown (*left*) are on a plane which has been warped by perspective effects in the imaging process. By computing the inverse transform it is possible to undo the effect of perspective (*right*).

image plane is parallel to the ground-plane. This is done by computing the planar projective transformation which is a linear transformation on homogeneous 3-vectors represented by a non-singular 3×3 matrix. For details see [65].

The easiest way to compute the projective transform required to rectify an image is to select, in the image, a set of points corresponding to a planar section of the world. Image coordinates and world coordinates are selected as shown in Figure 3.11.

It is important to note that the rectification achieved in this way does not require any knowledge of the camera's parameters or the pose of the plane. We show the effect of this on a full frame in Figure 3.11. However we do not want to compute the entire frame's projection, just the gaze so that we can determine what can really be seen in the world by the person. This is demonstrated in Figure 3.13.

Correcting gaze angles under perspective imaging

In order to display where the person is looking in the images angles are assigned to the discretised head-poses shown in Figure 3.1 according to the "compass" e.g. $N : 0^o$ etc. However, when the field-of-view is superimposed on the image (and, more importantly, when visibility of other

objects in the scene is determined using this field-of-view) it is important to correct for the fact that the camera is not fronto-parallel to the scene as for the acquisition of training data.

The angles are then corrected for the projection of the camera at each time step depending on the location of the person on the ground-plane in the image.

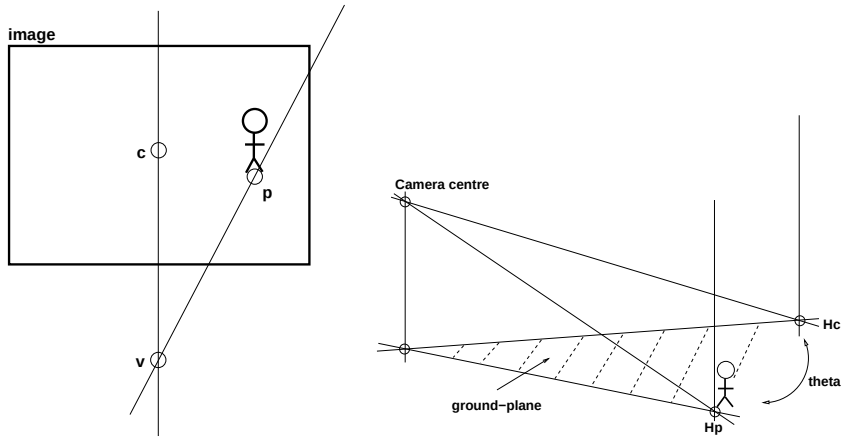


Figure 3.12: When assigning angles to the matched discretised head-poses one must compensate for the camera projection since "North" (see Figure 3.1) does not in general correspond to vertical in the image plane.

In order to choose the correct frame of reference we do not perform full camera calibration but compute the projective transform (**H** : image→ground-plane) by hand-selecting 4 points in the image as described above and shown in Figure 3.12. The vertical vanishing point, (**v**), is computed from the manual selection in the image of 2 lines which are known to be normal to the ground plane and parallel in the world. (See [65] §8.6 for details on how this relates to the "footprint" of the camera on the reference ground-plane). The angle θ between the projection of the optic-rays through the camera centre, **Hv**, and the image centre, **Hc**, and the point at the feet of the tracked person, **Hp**, is the angle which adjusts vertical in the image to "North" in our ground plane reference frame i.e.

$$\theta = \cos^{-1}[(\mathbf{Hc} \times \mathbf{Hv}).(\mathbf{Hv} \times \mathbf{Hp})] \tag{3.11}$$

(a) (b) (c)

Figure 3.13: Progression of improvements for visualising the gaze estimate: (a) No projection of the gaze onto the ground-plane and no compensation of the gaze angle (relative to the camera-centred frame) is used to generate this image, (b) In this image, gaze is projected onto ground-plane but perspective alterations in the assigned angle are not computed, (c) Gaze angle is computed using the projection from camera-frame to world-frame to create the final estimate of what the person can see.

3.4 Gaze estimation

3.4.1 Bayesian fusion of head-pose and direction

The naive assumption that direction of motion information is a good guide to what a person can see has been used in Figure 3.15. However, it is clear the crucial interaction between the two people is missed. To address this issue we compute the joint posterior distribution over direction of motion and head pose. The priors on these are initially uniform for direction of motion, reflecting the fact that for these purposes there is no preference for any particular direction in the scene, and for head pose a centred, weighted function that models a strong preference for looking forwards rather than sideways. The prior on gaze is defined using a table which lists expected (i.e. physically possible) gazes and unexpected (i.e. non-physical) gazes.

We define g as the measurement of head-pose, d is the measurement of body motion direction, G is the true gaze direction and B is the true body direction, with all quantities referred to the ground centre. We compute the joint probability of true body pose and true gaze:

$$P(B, G|d, g) \propto P(d, g|B, G)P(B, G) \tag{3.12}$$

Now given that the measurement of direction d is independent of both true and measured gaze

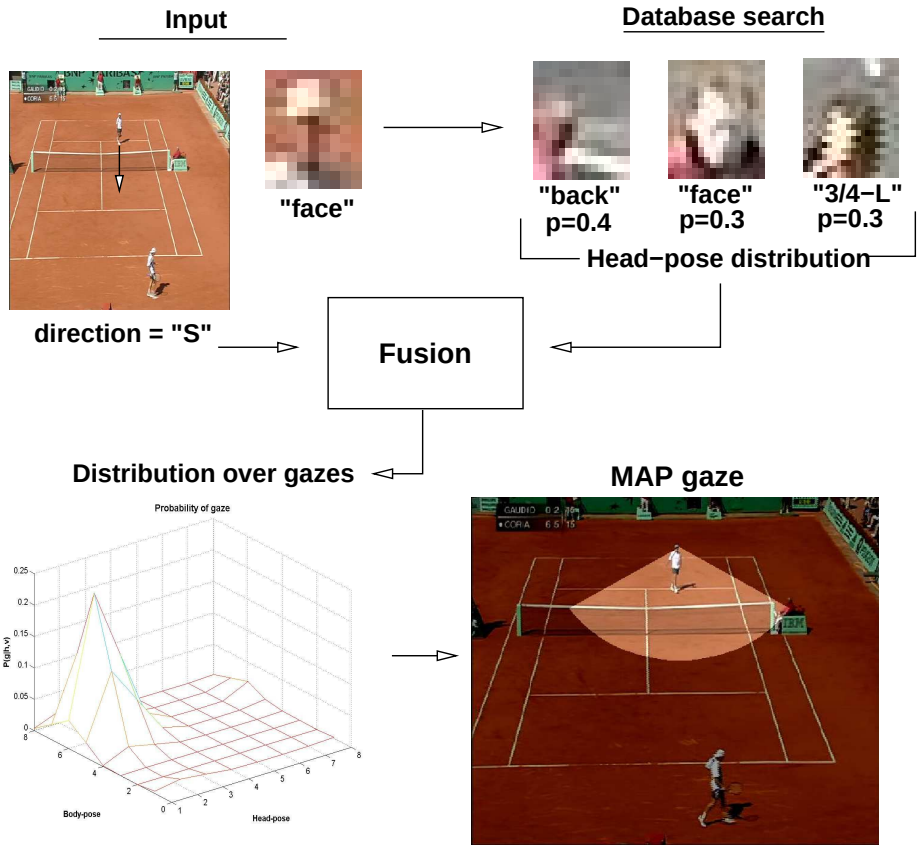


Figure 3.14: Fusing head-pose and direction estimates improves gaze estimation. Here, the ML match for head pose would be incorrectly chosen as "back" (*top-right*). The body-direction is identified as "S" (*top-left*). Since it is not possible to turn the head through 180^o relative to the body the resultant gaze from the direct combination of "back" and "S" has a low (predefined) prior. It is rejected as the most likely at the fusion stage only because of the anatomical constraints which are encoded in the priors. The MAP gaze is identified as "Face" which is a very good approximation to the true gaze.

G, g once true body B pose is known,

$$P(d|B, G, g) = P(d|B) \tag{3.13}$$

Similarly the measurement of gaze g is independent of true body pose B given true gaze G, i.e.

$$P(g|B, G) = p(g|G) \tag{3.14}$$

Then we have

$$P(B, G|d, g) \propto P(g|G)P(d|B)P(G|B)P(B) \tag{3.15}$$

We assume that the measurement errors in gaze and direction are unbiased and normally distributed around the respective true values

$$P(g|G) = \mathcal{N}(G, \sigma_G^2), P(d|B) = \mathcal{N}(B, \sigma_B^2) \tag{3.16}$$

(actually, since these are discrete variables we use a discrete approximation).

The joint prior, $P(B, G)$ is factored as above into

$$P(B, G) = P(G|B)P(B) \tag{3.17}$$

where the first term encodes our knowledge that people tend to look straight ahead (so the distribution $P(G|B)$ is peaked around B, while $P(B)$ is taken to be uniform, encoding our belief that all directions of body pose are equally likely, although this is easily changed: for example in tennis one player is expected to be predominantly facing the camera).

While for single frame estimation this formulation fuses our measurements with prior beliefs, when analysing video data we can further impose smoothness constraints to encode temporal

coherence: the joint prior at time t is in this case taken to be

$$P(G_t, B_t|G_{t-1}, B_{t-1}) = P(G_t|B_t, B_{t-1}, G_{t-1})P(B_t|B_{t-1}) \tag{3.18}$$

where we have used an assumption that the current direction is independent of previous gaze[3], and current gaze depends only on current pose and previous gaze. The former term, $P(G_t|B_t, B_{t-1}, G_{t-1})$, strikes a balance between between our belief that people tend to look where they are going, and temporal consistency of gaze via a mixture i.e.

$$G_t \sim \alpha\mathcal{N}(G_{t-1}, \sigma_G^2) + (1-\alpha)\mathcal{N}(B_t, \sigma_B^2) \tag{3.19}$$

Now we compute the joint distribution for all 64 possible gazes resulting from possible combinations of 8 head poses and 8 directions. This posterior distribution allows us to maintain probabilistic estimates without committing to a defined gaze which will be advantageous for further reasoning about overall scene behaviour. Immediately though we can see that gazes which we consider very unlikely given our prior knowledge of human biomechanics (since the head cannot turn beyond 90^o relative to the torso [109]) can be rejected in addition to the obvious benefit that the quality of lower-level match can be incorporated in a mathematically sound way. An illustrative example is shown in Figure 3.14.

3.5 Results

We have tested this method on various datasets (see Figures 3.15, 3.16, 3.17, 3.18 and 3.19).

These experiments can be categorised as follows:

- Detecting interactions that would be missed without knowledge of gaze independent of body direction (Figures 3.15 and 3.16).

[3] Although we do recognise that this may in fact be a poor assumption in some cases since people may change their motion or pose in response to observing something interesting while gazing around

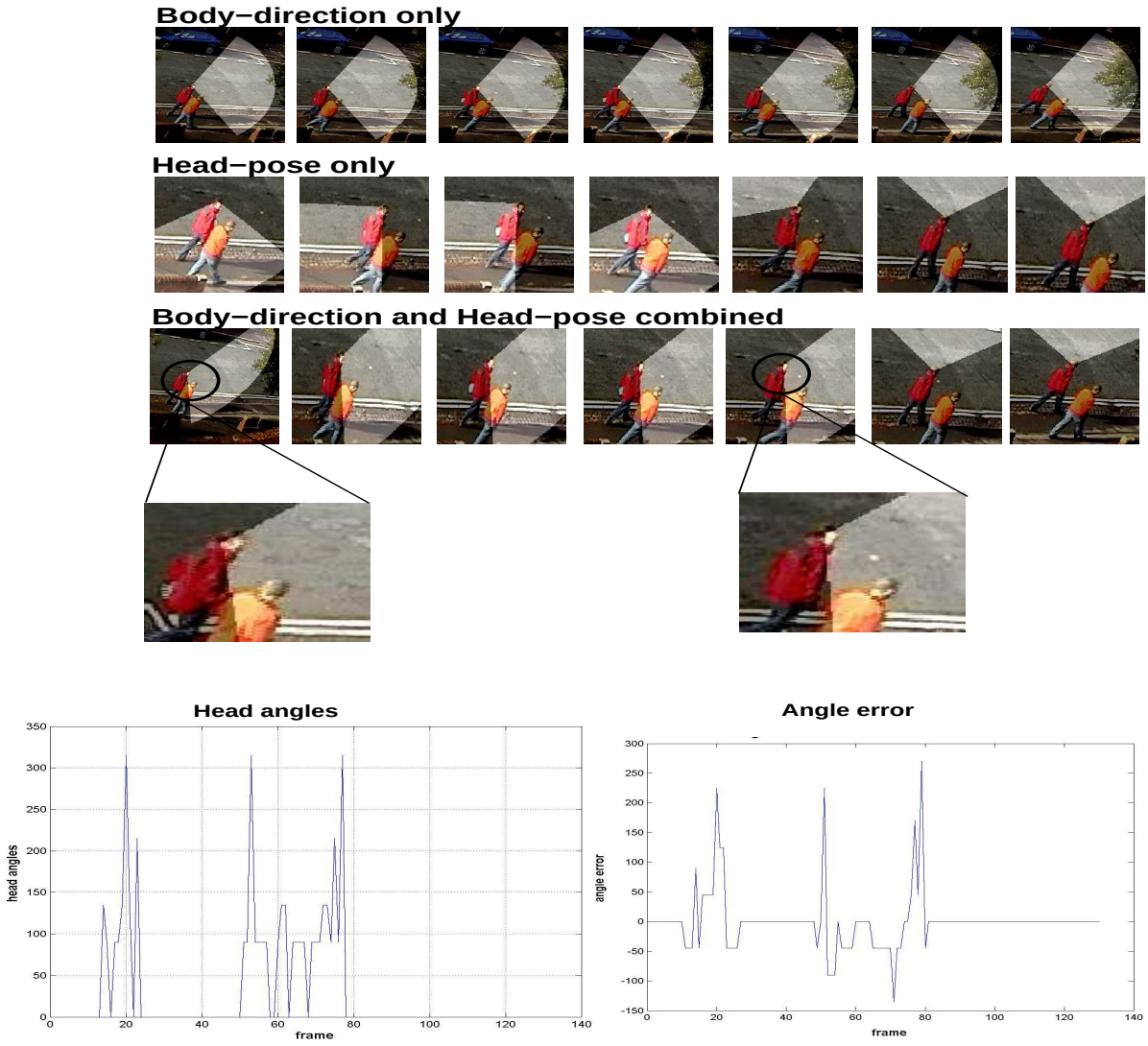


Figure 3.15: In this video there is an interaction between two people. The fact that they look at each other is the prime indicator that they are "together". On the first row we estimate gaze from body direction alone. On the second row gaze is estimated using head-pose alone, which gives an improved result, as far as detecting the interaction is concerned, but this is still prone to some errors. We see that (*third row*) fusing the head-pose and body-direction estimates gives a significantly improved result when it is the interaction that is required to be identified. That is, the "head angles" graph clearly shows two main head-turning events, the first short, the second longer. The angle-error is computed by comparing the estimated head-angles to hand-labelled ground-truth.

Figure 3.16: Two people meeting could potentially be identified by each person being in the other's gaze (in addition to other cues such as proximity), as we show in this example.

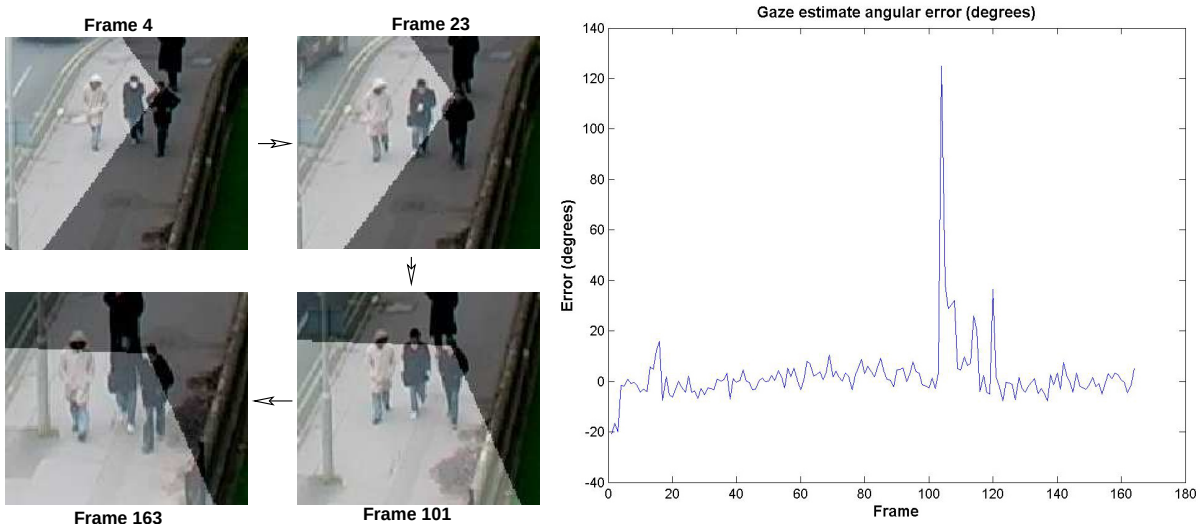


Figure 3.17: Second surveillance sequence. The same training data set as used to obtain the results above is used to infer head pose in this video without temporal smoothing. The ground truth has been produced by a human user drawing the line-of-sight on the images, quantised to 1^o. The mean error is 5.64^o, the median 0.5^o. Note that this is low due to the correct gaze falling very near a quantised value and is not necessarily representative of the general case.

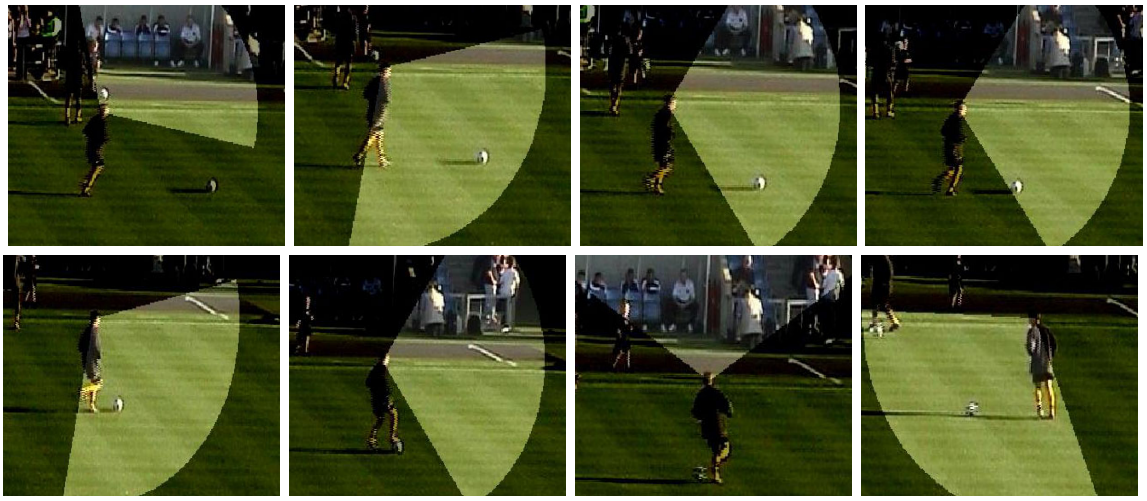


Figure 3.18: This example demonstrates the method in soccer footage. The skin histogram is defined only at the start of this sequence to compensate for lighting changes, but the exemplar database remains the same as that constructed initially and used on all the sequences i.e. it contains no examples from this sequence.

- Standard surveillance sequences which are used for evaluation (Figures 3.17 and 3.19).
- Sport (Figure 3.18), e.g. soccer, tennis.
- Failure modes (Figure 3.20).

3.5.1 Detecting interactions

The first dataset provided us with the exemplar data for use on all the test videos shown in this chapter. In the first example in Figure 3.15 we show significant improvement over using head-pose or direction alone to compute gaze. The crucial interaction which conveys the information that the people in the scene are together is the frequent turning of the head to look at each other. We reliably detect this interaction as can be seen from the images and the estimated head angle relative to vertical.

3.5.2 Surveillance sequences

The second example is similar but in completely different scenes from the training data. The skin histogram is recomputed for this video but the training data remains the same as for all the examples shown here. Once more the interaction implied by the head turning to look at his

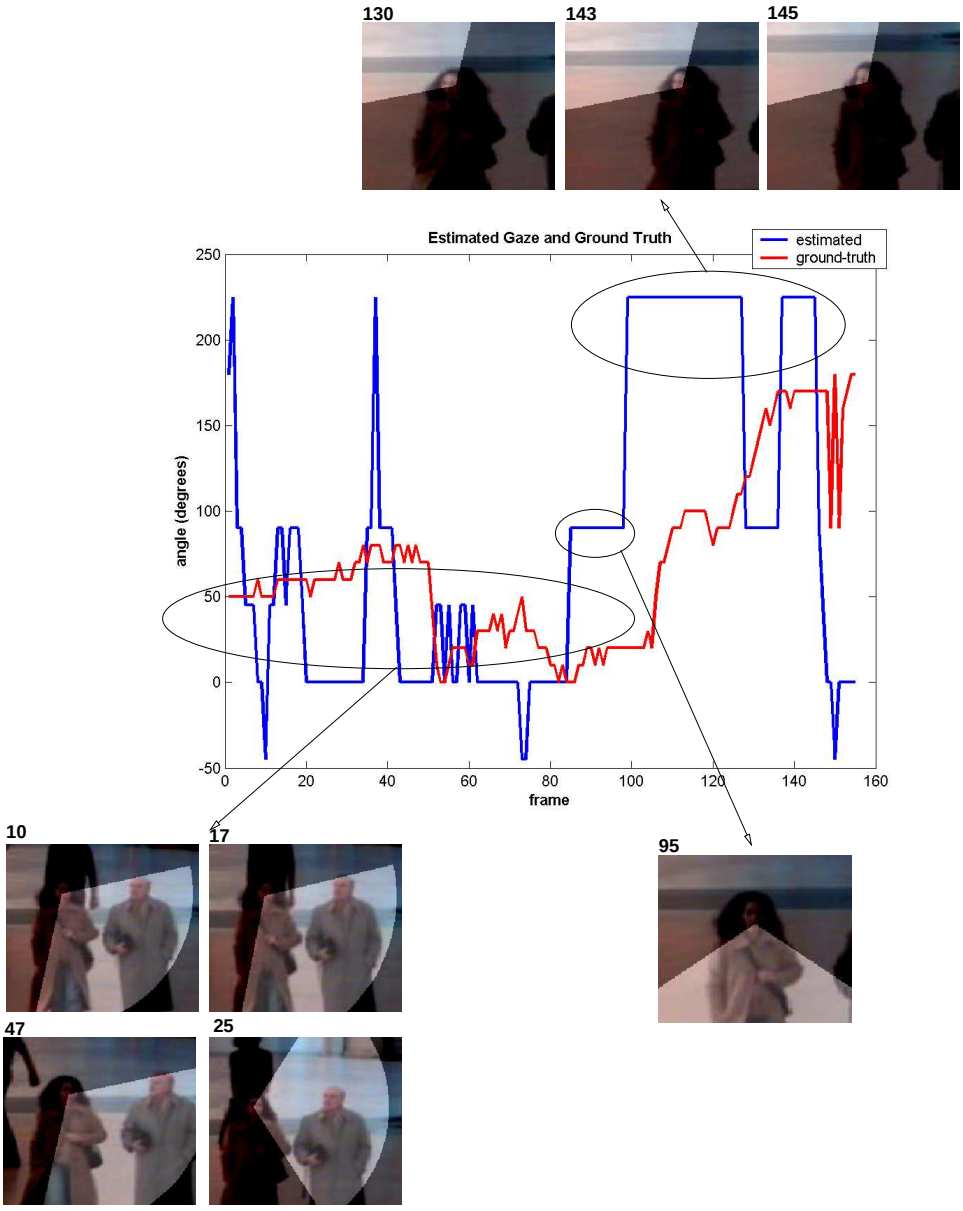


Figure 3.19: This figure shows the method tested on a standard sequence (see http://groups.inf.ed.ac.uk/vision/CAVIAR/). The errors are exacerbated by our discretisation of gaze (accurate to 45^o) compared to the non-discretised ground truth (computed to 10^o from a hand-drawn estimate of line-of-sight which we take to be the best-estimate a human can make from low-resolution images) and tend to be isolated (the median error is 5.5^o). In most circumstances it is more important that the significant head-turnings are identified, which they are here, as evidenced by the expanded frames.

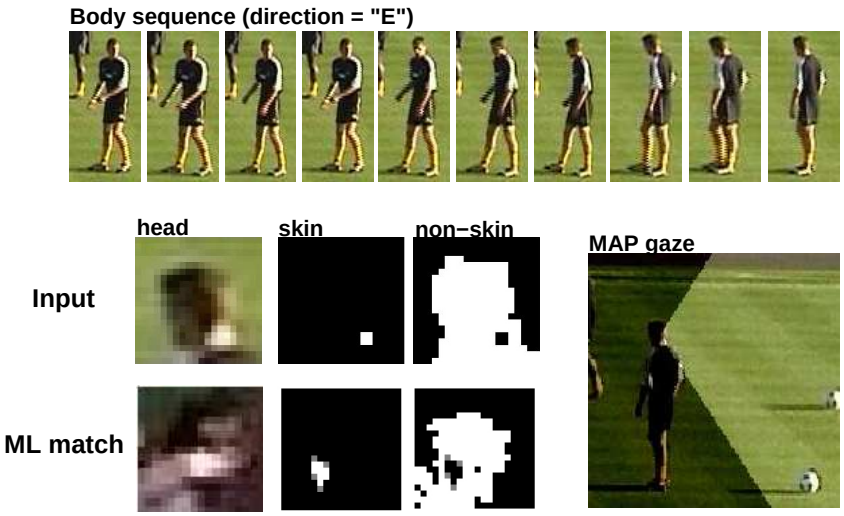


Figure 3.20: We show an example here where our method can fail. The mean body direction of the player (in the frames prior to the frame for which we estimate the gaze) is East, since he is moving backwards as his head rotates. The ML match is clearly not correct because the neck has been detected and there is no representation of gaze where the neck is visible in the training dataset. Fusing the direction and head-pose estimate results in the MAP gaze "side-LR", as expected, but incorrect. The reasons for failure are clear: body direction is not a good guide to gaze in this case and there is an unusual input which results in an incorrect match. Therefore, the predefined weights on gaze-direction given certain head-poses cause a better estimate of gaze to be rejected, in this case. Note that by adjusting the prior weights this condition can be relaxed.

companions is detected. The ground truth used to produce the graph in Figure 3.17 is obtained by drawing a line in the image corresponding to the best estimate an expert can make. The angle is computed to the nearest 10^o since it is unlikely an estimate at this scale can be more accurate.

The method is also tested on a standard vision sequence which has hand-annotated ground-truth data (which can be found at `http://groups.inf.ed.ac.uk/vision/CAVIAR/`). The results and comparison to ground-truth is shown in Figure 3.19.

3.5.3 Sports footage

We demonstrate the method on sports video in Figure 3.18 . It is shown in Figure 3.16 how useful this technique can be in a causal-reasoning context where we identify two people looking at one another prior to meeting.

3.5.4 Failure modes

In Figure 3.20 we show a combination of conditions unfavourable to estimating gaze accurately using the method we have described in this chapter. Specifically these are:

- Unmodelled skin. This is where skin not relating to the face region has been detected and the resulting head-pose estimate is therefore inaccurate.
- Unusual body direction. This can compound the error introduced by the unmodelled skin. For example a person walking backwards while looking forwards is unusual.

Note that there are other scenarios where the assumptions we have made in our formulation could lead to imperfect gaze-direction estimation. For example, where someone is walking and looking consistently in a direction perpendicular to the direction of travel, which may occur in sport. In that case, we would propose that other contextual information such as silhouette may be useful (and indeed the Bayesian approach is naturally extensible) to aid disambiguation.

3.6 Conclusion

In this chapter we have demonstrated that descriptors, readily computed from medium-scale video, can be used robustly to estimate head pose. In order to speed up non-parametric matching into an exemplar database and to maintain probabilistic estimates throughout we employed a fast pseudo-probabilistic binary search based on Principal Components. To resolve ambiguity, improve matching and reject known implausible gaze estimates we used an application of Bayes' Rule to fuse priors on direction-of-motion and head-pose, evidence from our exemplar-matching algorithm and priors on gaze (which we specified in advance). We demonstrated on a number of different datasets that this gives acceptable gaze estimation for people being tracked at a distance.

3.6.1 Further work on this topic

The Bayesian fusion method we have used in this work could be readily extended to include other contextual data. We used body direction in this work but information such as the silhouette may be potentially useful in providing body-pose context for gaze direction inference, conditional on reliable silhouette segmentation in surveillance footage being demonstrated. Moreover, the descriptor for head-pose could be extended to include information from multiple cameras.

3.6.2 Comments

One source of error is the video tracker which can produce inconsistency in the positions of the skin pixels in the target frame. Matches are, to some degree, dependent on the location of the skin pixels in the centre of the frame and tracking inconsistency can cause discrepancies to arise. This needs to be investigated. Additionally a uniform skin-colour histogram would improve our method by preventing the re-initialisation of skin colour for different lighting conditions.

We commented at the start of the chapter that skin detection across different lighting conditions and faces is not a solved problem. In practise, we re-initialised the skin-colour histogram for new videos since the descriptor we chose was based explicitly on skin pixels and their relation to the size of the head and the non-skin pixels. In theory, it is possible that the descriptor not be based on skin *per se* but on homogeneous regions which can be identified as skin or non-skin. So if a 60^o view of a head is available it is clear, even at a low-resolution, that regions of pixels lying in a range of colours correlate with face/non-face regions of the head. It does not matter what exact range of colours are provided to represent the face and non-face histograms provided they are suitably distinct. Therefore a descriptor could be defined not on the basis of a universal skin histogram (or even one local to that video) but on the basis of distinct colour regions.

The novel method described here would be most useful in a causal reasoning context where knowledge of where a person is looking can help solve interesting questions such as, “Is person A *following* person B?” or determine that person C looked right because a moving object

entered his field-of-view. This is a topic we address in chapter 6 when lower-level techniques have been further developed in the chapters which immediately follow.

4

Action recognition

In this chapter we develop a system for recognising human action and for deriving commentaries on activity by giving these actions spatial context. Actions are described by a feature vector comprising both trajectory information (position and velocity), and a set of local motion descriptors which represent person-centred motion (walking etc.). Action recognition is achieved via probabilistic search of image feature databases representing previously seen actions. Human actions are represented using a hierarchy of abstraction: from actions centred on the person, to actions with spatio-temporal context, to action sequences. While the upper levels all use Bayes networks and belief propagation, the lowest level uses non-parametric sampling from a previously learned database of actions. The combined method represents a general framework for human activity recognition. We demonstrate the results on broadcast tennis sequences and urban surveillance footage for automated video annotation.

The work described in this chapter was published in the proceedings of the International Conference on Computer Vision, Beijing, 2005 [131] and has been submitted to the journal Computer Vision and Image Understanding [133].

4.1 Introduction

A system capable of inferring the behaviour of humans would have many applications from visual surveillance in the military and civilian domains to automatic sports commentary. In particular, a method for classifying an instantaneous human action, or even better, determining a behaviour that may comprise several actions in sequence, would inevitably be a core building block of such a system. In this paper we present progress towards such a system by demonstrating how a non-parametric learning and classification technique for actions, can be combined with an effective, parametric representation of action sequences, which we use to describe behaviours.

The lowest level of our system, for recognising actions (e.g. *walking* versus *running*, versus *standing*) is based on the technique described by Efros *et al.* [44] who showed how action recognition can be structured as a search over a comprehensive training database. Though their work was effective for matching frames in video sequences according to similar gross properties of inter-frame motion, the instantaneous action descriptors used are only effective if the training set is very large indeed. In many applications, including our own, there is a need to achieve similar recognition rates but with a much smaller training set. To this end we show how an extension to their "blurry motion channel" descriptor can effectively disambiguate between types of action even though the intra-sequence description of each frame of different actions are very similar.

Efros *et al.* deliberately used position independent descriptors, and made no attempt to reason at a higher level about the actions. We are explicitly interested in higher-level reasoning about action context. In particular, the spatial context (where an action happened) and the temporal context (when it happened, and more interestingly, where it occurred in a sequence of actions) are vital for higher level reasoning and thus we take steps to represent both. For example an action "standing still" may be interpreted as normal behaviour in one spatial context (at a bus stop e.g.), while it may be considered to be the higher-level behaviour "loitering" if it occurs in an alleyway. To this end, we consider position and velocity information as additional features; these too are compared against a training database to elicit (respectively) qualitative position

and direction labels. In an urban surveillance scenario these qualitative descriptors might be, for example, *nearside-pavement*, *on the road*, *far-side pavement* for position, *left-to-right*, *away*, *towards* (etc.) for direction. The results of the three database searches are then fused using a Bayes net to provide a distribution over possible *spatio-temporal actions* (an example of a spatio-temporal action might be *walking, left-to-right, near-side pavement*). Taking the maximum likelihood (ML) spatio-temporal action at each instant in a sequence enables the system to construct a commentary of the (estimated) observed activity.

4.2 Chapter structure

Note that much of the literature relevant to this chapter has been reviewed in the main literature review presented in chapter 2. The remainder of this chapter is structured as follows. We begin with a more detailed description of each of the stages of our algorithm. Section 4.3 deals with the low-level non-parametric action recognition stage, and describes in particular how we have implemented an efficient probabilistic search of an exemplar training database in order to sample from the action (and qualitative position and direction) distribution(s). Section 4.3.4 describes the Bayes Networks that fuse the low-level data. Smoothing of the action sequences and inference of high-level behaviour is the subject of chapter 5). Section 4.4 gives experimental results and we conclude in section 4.5.

Throughout the chapter we use sequences from urban surveillance scenarios or sports footage. In our examples we assume the urban data represents one of a small set of actions such as *walking, running, standing, dithering* and a reasonable range of qualitative positions i.e. *nearside-pavement, road, driveway, farside-pavement* and directions e.g. *left-to-right, across* etc. This set of sequences is used to test the action matching and spatio-temporal action recognition steps. A richer set of actions is found in tennis. Using our method we show that an intermediate representation of action can provide an automatic commentary. (This commentary can be improved by smoothing the action sequences using an HMM which encodes expert knowledge about shot transitions e.g. that a serve starts a point and that a non-shot, such as *running*, follows a shot, as we show later in chapter 5).

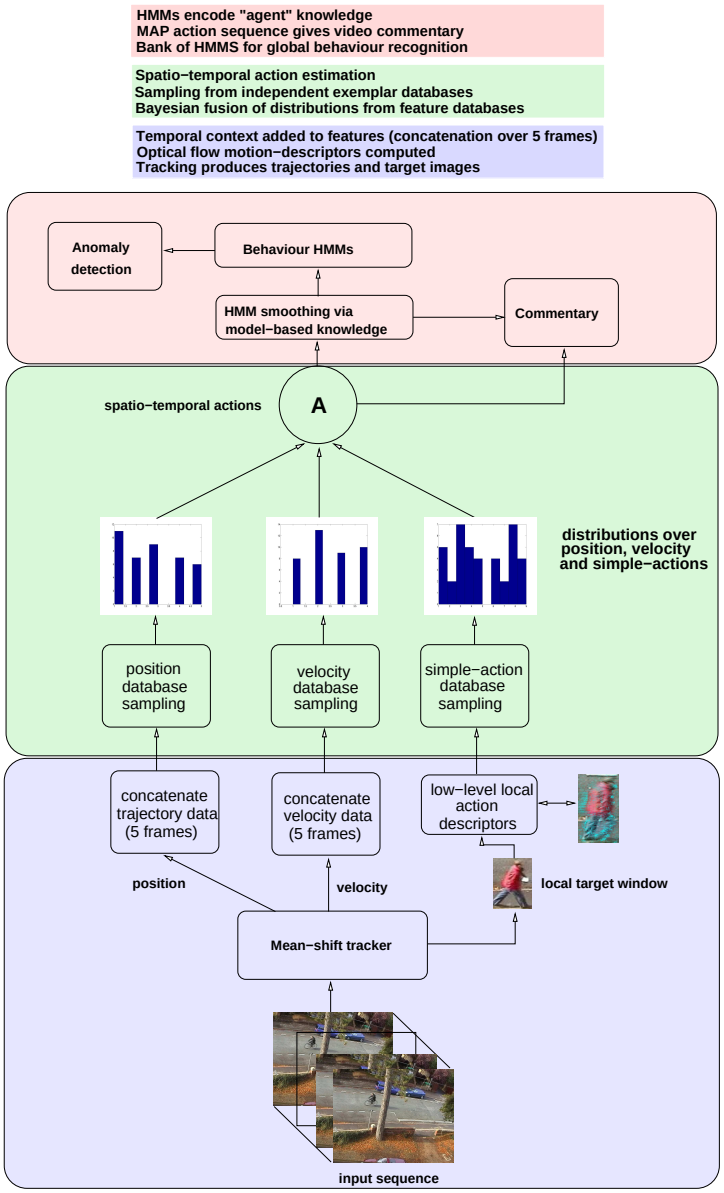


Figure 4.1: This schematic diagram illustrates the relationship between image features, actions, action sequences and the high-level parameterisation of behaviour which are described in chapters 4 and 5. Databases of the position, velocity and motion-descriptor features are prepared in advance and are hand-labelled with qualitative descriptions of place, direction and action. Distributions over each of these features are computed via non-parametric sampling of the databases. These distributions are combined using a Bayes Net which produces a distribution over spatio-temporal actions. This provides a text commentary of observed activity. Sequences of actions are also encoded as HMMs allowing higher-level descriptions of overall activity to be inferred. These HMMs are encoded using the spatio-temporal actions and not directly from image data.

Figure 4.2: Fixating on a target using a colour-based tracker. The extracted target image is shown in the expanded images along various points of the target centroid trajectory, showing tracking successfully in scale and image-space.

4.3 Action recognition

The main component of our instantaneous action recognition method is action recognition via non-parametric matching of trajectory data and instantaneous motion descriptors, fused via a Bayes net.

4.3.1 Target description

Using a mean shift tracking algorithm [31] (described in detail in appendix A), we extract the following information for each target for each frame: position, velocity and a window around the target (see Figure 4.1). In addition to the target's place and speed we are also interested in classifying the action of the person we have tracked e.g. *walking* or *running*. An effective method to do this was suggested by Efros *et al* [44]. In that work, a local motion descriptor based on coarse optic flow is extracted from a stabilised target window. (This pre-processing step was

performed in chapter 3, see Figure 3.2.) This local motion descriptor is compared against a dataset of previously seen local motion descriptors that have been hand-labelled with their corresponding actions. The nearest-neighbour match provides an action label for the current data. The optic flow between consecutive frames of a sequence is computed[1]. The optic flow vector-field $\mathbf{F}$ is split into two scalar fields which are the horizontal and vertical components of the optic flow field, F_x and F_y. These are then half-wave rectified into positive channels F_x^-, F_x^+, F_y^- and F_y^+ such that:

$$F_x = F_x^+ - F_x^- \tag{4.1}$$

$$F_y = F_y^+ - F_y^- \tag{4.2}$$

Each of the channels is blurred with a Gaussian kernel and normalised, producing the four motion descriptors for every frame of the sequence $\hat{F}b_x^+$, $\hat{F}b_x^-$, $\hat{F}b_y^+$ and $\hat{F}b_y^-$.

A version of normalised cross-correlation is further employed such that, if the four motion-channels for frame i of a sequence A are defined to be a_1^i, a_2^i, a_3^i and a_4^i (similarly for frame j of the sequence B), then the similarity between motion descriptors centred at frames i and j is given by:

$$S(i,j) = \sum_{t\in T}\sum_{c=1}^{4}\sum_{x,y\in I} a_c^{i+t}(x,y)b_c^{j+t}(x,y) \tag{4.3}$$

and, when the matrix A_1 is defined as the concatenation of all a_1 vectors (similarly for the other channels, and for sequence B), the frame-to-frame similarity matrix between the two sequences is:

$$S = A_1^T B_1 + A_2^T B_2 + A_3^T B_3 + A_4^T B_4 \tag{4.4}$$

[1]Optic flow is ideal for this purpose because it is photometrically invariant and invariant to clothing or appearance [89]. Invariance is essential as we are seeking a general description of the incremental motion of a person to match the action between different "actors".

T is defined in the original work of Efros *et al.* as "the temporal extent of the descriptor". Although equation 4.3 implies that, by varying T, temporal context can be achieved, in practice, T is defined when the descriptor is computed, initially.

For frame-to-frame optic flow, therefore $T = 1$, or at most $T = 2$. It is not clear that, unless encoded in the descriptor itself, that Efros *et al.* intended this term to allow for temporal context in the descriptor, as this is not discussed in the original work [44].

Further, Efros *et al.* recognise that if the sequences A and B are similar but occur at different rates the similarity matrix will have strong responses along the off-diagonal elements and so S is convolved with a kernel which is a weighted-sum of near-diagonal lines:

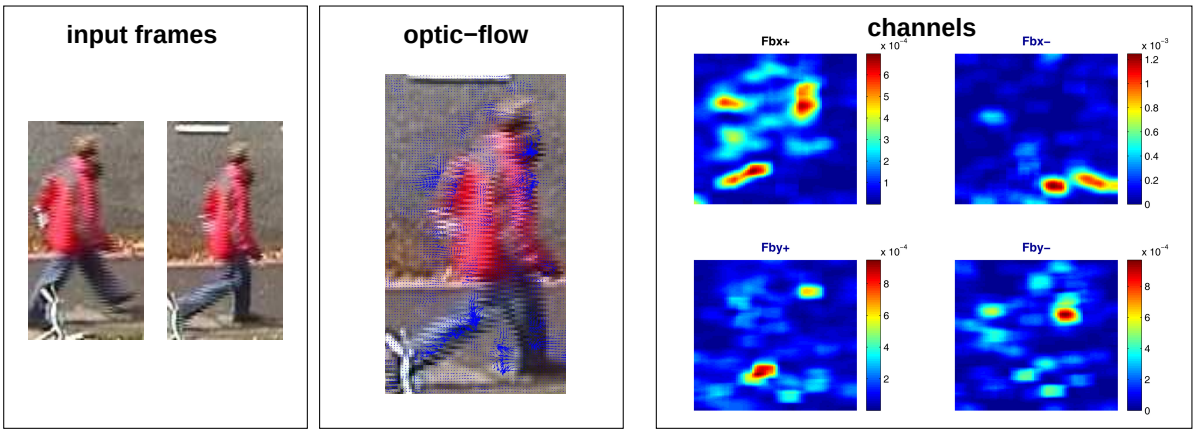
$$K(i,j) = \sum_{r \in R} w(r)\chi(i, rj) \tag{4.5}$$

where R is a range of rates.

4.3.2 Optic flow computation

Optic flow is a measure of image-velocity. In estimating optic flow the aim is to compute an approximation to the 2-D motion-field which is a projection of the 3-D velocities of surface points onto the image plane [156, 70]. There exist a number of methods for estimating the optic flow field. Barron *et al.* have reported on a comprehensive study of the most common methods in each of these categories [5]. While they do not conclude that one method is consistently superior than all others, it is apparent from the experiments performed that the Lucas and Kanade technique [89] is among the best for the quantitive experiments performed by Barron *et al.* [5]. The average error across synthetic and real sequences was reported as 1.06^o. Therefore, we employ the Lucas-Kanade algorithm (derived in appendix E) for computing optic flow for the following reasons:

1. Efros *et al.* used the Lucas-Kanade algorithm in "Recognising action at a distance" [44] and to construct a fair analysis of the performance of their method in our application



(a) (b) (c)

Figure 4.3: The optic flow is a measure of how pixel information is translated in an image between successive frames (i.e. 2-D image velocity). (a) In this example, the pair of input images are shown on the left. (b) The flow vectors are shown for the raw optic flow superimposed on one frame. (c) The Gaussian blurred optic flow in the x and y direction is further split into the four (blurred) non-negative channels described in the text. Combined, these channels comprise the descriptor of instantaneous action defined by Efros in [44] and used as the basis for the action-recognition stage in this work.

domains this aspect must remain unchanged,

2. It gives the best performance i.e. the angular error is proved to be the smallest of all common optic flow measures,

The results of the Lucas-Kanade method (the algorithm is derived in appendix E) applied to images of a person walking are shown in Figure 4.3.

The action-recognition method based on the Efros *et al.* optic flow descriptors (which are shown in Figure 4.3) works well only if the newly-observed sequence for which one wishes to find a best match is represented in the example set.

For every example sequence in the exemplar set (which can be regarded as a "database"), the best match can be found at any time step by using the similarity matrices of equation 4.4 as a lookup table. These matrices are shown in Figure 4.4 for a new sequence compared to four exemplar sequences.

If we consider a standard "surveillance" scenario, it is clear that there are typical patterns of motion. At a traffic intersection, for example, people and vehicles do not move entirely freely.

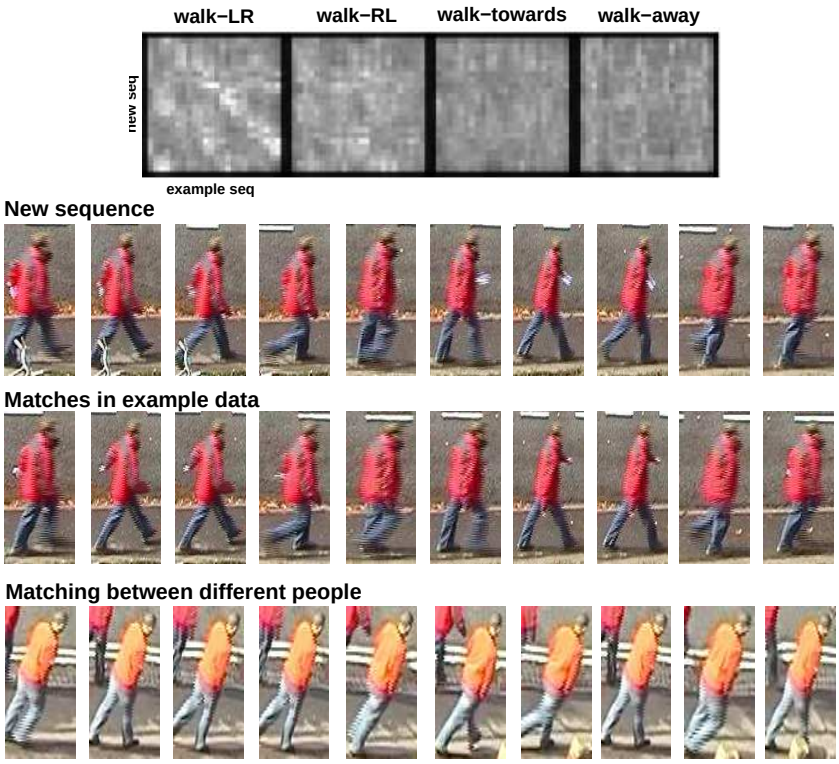


Figure 4.4: The action-recognition technique of Efros *et al.* is demonstrated here. Our example database comprises 4 sequences and the frame-to-frame similarity matrices are shown for each of these models given the new input sequence (shown in middle row). Note that for the best-matching *sequence* there is evidence of periodic structure in the similarity matrix (*left* "walk-LR"). The best matching frame in the database at each new input frame is chosen from the similarity matrices. The input frames are at the top and, directly below, the best matching frame is shown. (Note that the background clearly shows these are different frames from separate sequences. For completeness, in the *third row* we show matching is effective despite the fact that the appearance of each person is quite different.

Figure 4.5: A slight variation of normal activity is introduced in the new sequence (*top row*). Despite this being only subtly different from the normal activity modelled (non-parametrically) in the database, mismatches very quickly occur.

The pavements and roads dictate the trajectories of people in the world. A limited set of camera views is often available, which, taken with the constraint of the people in the scene, limits the appearance of people in the video considerably. This is a crucial point because, by definition, the majority of people act in a "normal" fashion. When subtle variations of this activity appear, there is a danger that the activity will be mis-labelled because that variation has not been captured in the training data. We show an example of this problem in Figure 4.5.

To obtain the experimental results shown in Figures 4.4 and 4.5 we captured four exemplar sequences from surveillance data which represent the typical motion in the scene, as decided by an expert user. By matching the same person at different times we minimise errors due to variations in size and shape (although the method can deal with these variations, as we see in Figures 4.6 and 4.7). The database is small, comprised of 200 frames of data from four motions. These are labelled by the user as "walking, left-to-right", "walking, right-to-left", "walking, away-from-camera" and "walking, towards-camera" and have corresponding short titles (used in the figures) `walk-LR`, `walk-RL`, `walk-away` and `walk-towards`.

The sensitivity of the Efros *et al.* technique to the lack of comprehensive data is revealed by the fact that significant mismatches occur for even subtle variations in motion. For example, Figure 4.5 shows that a "wandering" motion cannot be reliably correlated with "walking, left-to-right" despite this being the most obvious choice (to human eyes) whereas Figure 4.6 shows that the motion reverses.

An issue not addressed by Efros *et al.* is that of cluttered environments. The examples provided in [44] are mainly in the sports domain. This assumption does not translate to urban environments where frequently a person's limbs are obscured by immovable static objects (e.g. lamposts, trees) and by non-static objects (e.g. people, cars). As we have seen, this is likely to cause considerable problems for choosing the best-matching model *even if the exemplar data were comprehensive.*

A motivating factor for the approach we take, based on this technique, is the need to be able to model *how close* a particular match is to what has been seen before. This is particularly important given the fact that (a) training data will not necessarily be as complete as desired, (b) objects in the world cause obscuration or even occlusion. In order to solve this we propose a probabilistic solution to the instantaneous action recognition problem. Before that is introduced we describe the benefits of adding temporal context to the motion descriptor.

4.3.3 The significance of temporal context in the descriptor

In our experiments we have found that if the database contains only a small number of examples of a certain action the risk of the nearest-neighbour being incorrect is greatly increased. In order to add temporal context and mitigate against this type of confusion, we create a richer feature descriptor by concatenating the coarse motion descriptors from a number of consecutive frames, typically 5, to form a motion feature vector at each frame. An example showing the benefits of this enhancement is shown in Figure 4.6. Efros *et al* deliberately discarded all positional information. In contrast we have argued in section 1 that such information is important in placing an action in its spatial context. To that end we also create additional databases of previously seen trajectories (position and velocity). In each case the feature vector is the concatenation of a few (typically five) frames worth of position (respectively velocity) data, and the database examplars are labelled with qualitative position (respectively, qualitative direction) labels. The databases of position, velocity and local motion are maintained independently, and the set of "normal" actions is the set of combinations of the qualitative labels attached to the exemplars in the feature databases. Matches from the position, velocity and motion-descriptor

databases are fused using a Bayes net described in Section 2.3. Prior to that, we discuss the database organisation and search techniques. This is not trivial for two reasons (i) the volume of data from the blurry motion descriptors presents a challenge for efficient search: there are 30000 entries in a single local motion feature vector for a 30×50 pixel target; (ii) for more effective data fusion (and necessarily for appropriate use of a Bayes net) we do not simply want a nearest-neighbour (i.e. maximum likelihood) match, but rather a distribution over possible matches.

Using the efficient sampling technique which we described in chapter 3 the search time is improved by a factor of 20 and, since we sample many times, the search provides a set of particles which represents a distribution over matches of position, velocity and motion-descriptor into frames of the previously seen examples. An example of such a distribution is shown in Figure 4.7. The database was created using 60 minutes of automatically tracked (but hand-labelled) data, and was tested using novel sequences of similar actions.

4.3.4 Action likelihood computation

Complex motions are composed of primitives. We define a "simple-action" as a target-centred action such as *walking*. This can be estimated by sampling from the motion-descriptor database alone. By fusing the likelihoods of the matches from the position, velocity and motion-descriptor exemplars we compute the probability of a spatio-temporal action such as *walking-left-to-right-on-nearside-pavement*. We use a Bayes Net to effect this information fusion: if the spatio-temporal action is denoted a, x is the qualitative position, v is the qualitative direction, and m is the simple-action, then assuming conditional independence yields

$$p(a, x, v, m) = p(a)p(x|a)p(v|a)p(m|a) \tag{4.6}$$

The distributions $p(x_{match}|x_{input})$, $p(v_{match}|v_{input})$ and $p(m_{match}|m_{input})$ are estimated by sampling from the databases. We compute the marginal distribution $p(a)$ since, for any given data

Figure 4.6: Matching optic flow based motion descriptors without large volumes of representative data (i.e. a comprehensive data set representing the many nuances of particular actions) can result in incorrect matches, as shown here (*left*). In these examples walking in one direction has been confused with walking in the opposite. Instantaneously (i.e. over 2 frames) the movement of arms and legs is significant whereas the body direction is not. We concatenate the motion-descriptor data from 5 consecutive frames which provides temporal context and results in the Maximum Likelihood matching exemplar being less ambiguous (*left column*). The increase in data available for matching reduces the ambiguity which arises when the exemplar data may not be comprehensive, as in a surveillance scenario where people are predominantly viewed from a limited set of angles. The motion does not reverse in this case.

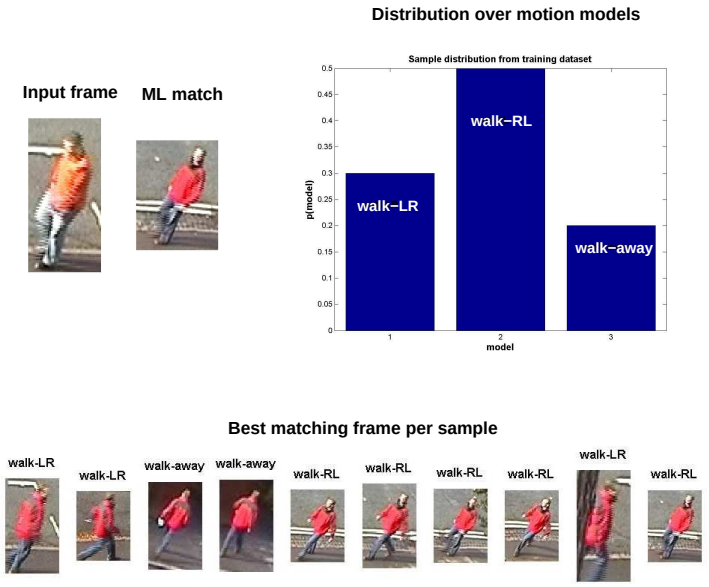


Figure 4.7: We illustrate the optic flow based action-recognition technique incorporated into a pseudo-probabilistic sampling from the exemplar database. The input frame (*top left*) is shown beside the ML frame from 10 samples of the motion-descriptor database. The more complete information is provided by the sampled distribution of matches from the database. These are shown *top right*: the distribution over model-types in the exemplar set, and, *bottom row*: the matching frames for each sample of the database.

Figure 4.8: This second scene (*left*) has a considerably richer set of actions. The example set is comprised of 27 different types of spatio-temporal activity with a range of person-centred actions from walking (in a variety of directions relative to the camera) to running and standing still, loitering etc. We show here a new example of the action walking matched into the exemplar database by taking the ML match from all samples at each frame. The input is on the top row, with the nearest matching exemplar frame directly beneath.

Sequence	**Total (frames)**	**Example database (frames)**	**Test sequences (frames)**
Urban street	5455	665	2361
Junction surveillance	76040	4491	18445
Tennis	90000	494	3132

Figure 4.9: The data volume for each of the videos used in the analysis of our technique described in this chapter.

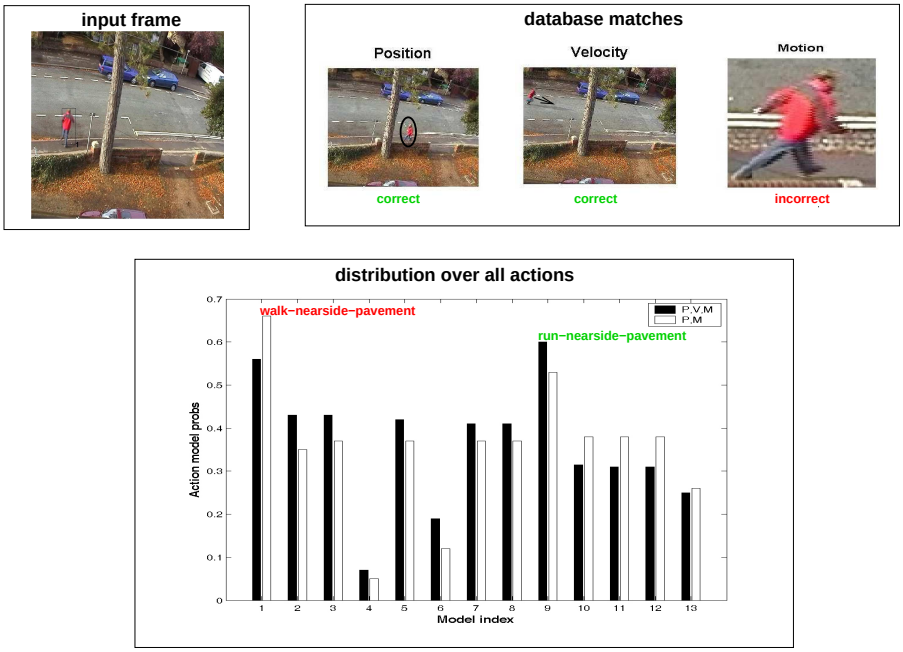


Figure 4.10: Velocity and motion-type are as important as position for action-recognition. Here the ML motion-type is (incorrectly) classified as *walking*. When the resulting distributions from each of the inputs (i.e. position, velocity and motion-type) are fused the ML estimate is now (correctly) *running-on-nearside-pavement*. The action probability distribution is shown here when velocity is excluded (white bars, red text) and included (black bars, green text).

d (here x, v and m),

$$p(d|a) = \frac{p(a|d)p(d)}{p(a)} \tag{4.7}$$

$p(a|d)$ is specified in the conditional probability table for the node a, $p(d)$ is defined from the frequency of occurrence of data d in the training set and choosing $p(a)$ to be uniform is suitable in most cases. Figures 4.11 and 4.10 illustrates this process for two different applications. Figure 4.10 highlights the significance of each input for successful action classification.

4.4 Automatic text commentaries of activity

One application of these techniques is the automatic generation of text commentary on observed activity. At each frame the distribution over all possible spatio-temporal actions is computed using the evidence from the position, velocity and simple-action recognition method described above. For a commentary the Maximum Likelihood action is chosen and the test description

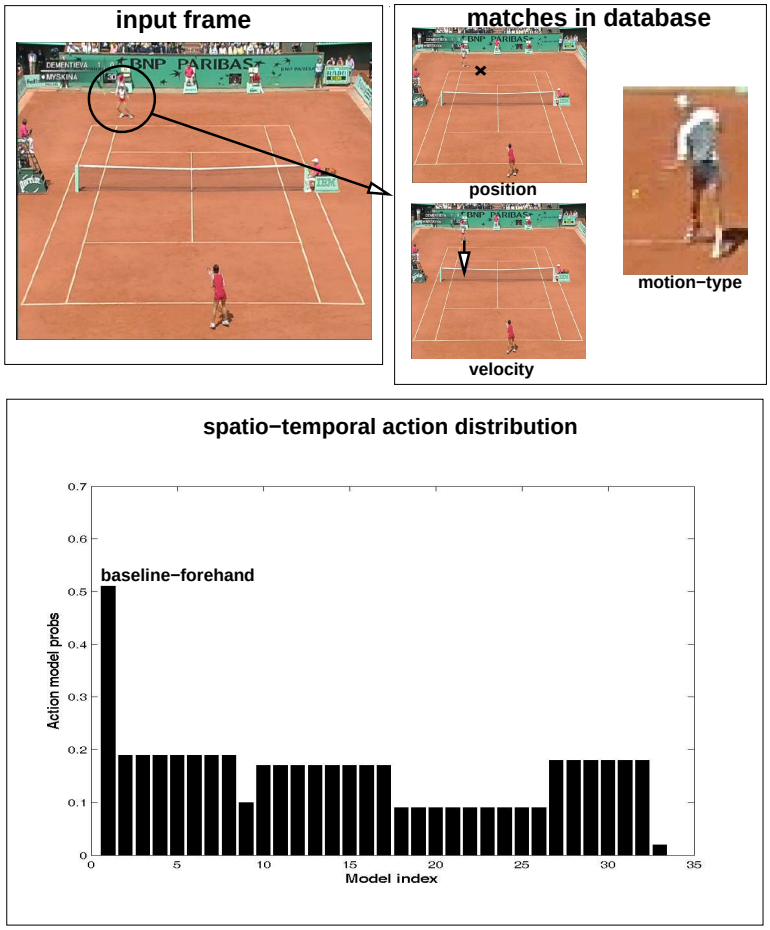


Figure 4.11: There are 33 possible shots resulting from combinations of positions and shot-types in our exemplar set. The closest ML matches in the databases for this frame are shown next to the still image in the order position, velocity and shot-type. The distribution over all shots is shown in the graph. The most likely shot is computed to be *baseline-forehand* which is correct.

of that spatio-temporal action is used to describe the person's activity at that instant. The validity or the accuracy of the description is dependent on (a) the descriptive language used to label the exemplar sequences in the databases of position, velocity and simple-action, (b) whether that action has been seen before.

The former dependence, (a) above, requires the expert user to ensure that the language used to described the scene is accurate. Errors in this regard would be a failing of the training phase, not the general approach (i.e. the "expert" is not as expert as believed).

The latter dependence, (b) above, is mitigated by the fact that each activity has a likelihood of occurrence and, as expected, even though the best match happened to be a certain spatio-temporal action, if that is, in fact a poor match overall (as interpreted by the user) then the likelihood reflects that.

4.4.1 Commentary of video from an urban location

For a real scene with a relatively small exemplar set (due to the fact that there is a limited set of typical activities) we demonstrate achieving a useful basic text commentary from the distribution of spatio-temporal actions in figure 4.12.

In Figure 4.13 an example of abnormal activity is captured. Because it is abnormal it is not represented in the exemplar databases of position or simple-action (although the velocity with the label *running* may be present). The resulting commentary therefore has a much lower likelihood of occurrence than the normal example of Figure 4.12. This likelihood can be interpreted as a measure of belief of the best-matching activity.

The second video is from a much more complex urban environment, one which is a real surveillance scenario. This elevation of the camera is typical of the placement of road monitoring systems (as anyone familiar with major roads in the U.K. can verify). In fact, as can be seen in Figure 4.17 this camera is viewing a traffic junction. This gives rise to a richer set of activity than the previous scene. It is also considerably more challenging in that objects are now in the medium to far field such as those shown in the first two frames of Figure 4.17.

Frame	Activity	Likelihood
1 - 70	Walking on far-side pavement	0.86
71 - 225	Walking on road	0.94
226 - 450	Walking on near-side pavement	0.94

Figure 4.12: An accurate commentary is obtained for this urban street scene where the person moving in from the top-right of the images (in orange) is under observation. The descriptions of activity are achieved using the labelling from the exemplar database.

Frame	Activity	Likelihood
1 - 75	Walking on near-side pavement	0.34

Figure 4.13: Neither the position nor the simple-action of the person in this (abnormal) sequence is well-represented by the predefined exemplar data, therefore the probability of the ML activity is significantly lower than that of Figure 4.12.

Initial foreground segmentation is performed by using a static background frame. This procedure is applied to each frame in the sequence and a sample of the effect of this process is shown in Figure 4.15.

Scene markup

The distinct regions in this scene are shown in Figure 4.14, with the labels attached to regions, possible activities and directions within this scene.

Some of the exemplar data is plotted in Figure 4.16. A professional surveillance officer has detailed knowledge of the area under observation at his disposal, and this knowledge is reflected in his reporting of the target behaviour. Therefore in Figure 4.16 the labels given to the pedestrian training examples we have shown here reflect, for example, that the road which runs vertical in the frames is known to run North-South. Moreover when the labels applied to the regions in which *drivers* are active, we see that the street names are applied[2]. The labels, as decided by the human user, are shown in the legend of the figure.

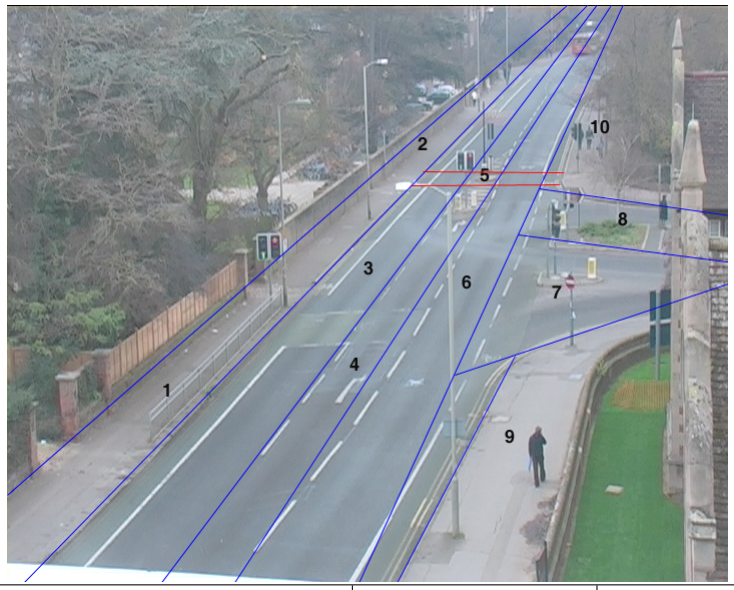
Automatic labelling of activity

By using our instantaneous action recognition technique we can derive text "explanations" of the observed activity of individuals using the exemplar data. We show the types of explanations achieved in Figure 4.17 above a representative frame where that action is found.

We further show how more complex activity can be recognised and described. In Figure 4.18 the example would correctly be described as "walking across the road at the traffic lights". Our automatic description is more exact, however, describing the location of the crossing and the exact location of the pavement ("NE-pavement") according to the labels provided by the user/trainer.

An additional example is that of "jogging across the road", shown in Figure 4.19.

[2]The expert in this case was an Oxford student!



Regions	Person actions	Directions
Northbound Lane (3)	Walking	North
Right Turn Lane (4)	Running	South
Southbound Lane (6)	Stopped	East
Parks Road Westbound (7)		West
Parks Road Eastbound (8)		Stopped
South-East pavement (9)		
South-West pavement (1)		
North-East pavement (10)		
North-West pavement (2)		
North pedestrian crossing (5)		

Figure 4.14: The scene is divided into regions and labelled by an expert analyst. The labelled regions, activities and directions for this scene are detailed in this table.

4.4.2 Commentary of tennis matches

We apply the technique to tennis video in order to classify each players' shots and produce an automatic text commentary. Following the tracking of players in video of 4 different professional tennis matches, we manually segmented the sequences into exemplars of standard tennis shots and created independent databases of the position, velocity and simple-action motion descriptors. The shots we extract exemplars for are labelled with the following qualitative descriptions: *forehand*, *backhand*, *forehand-volley*, *backhand-volley*, *serve*, *smash*. In addition we provide examples of non-shots labelled *running*, *walking* and *waiting-for-serve*. Shot example databases are created for each player i.e. facing the camera (farside court) and facing away from

Figure 4.15: (*Left*) Background subtraction on one entire frame and (*right*) the foreground segmentation of a person extracted from a similar background-subtracted image.

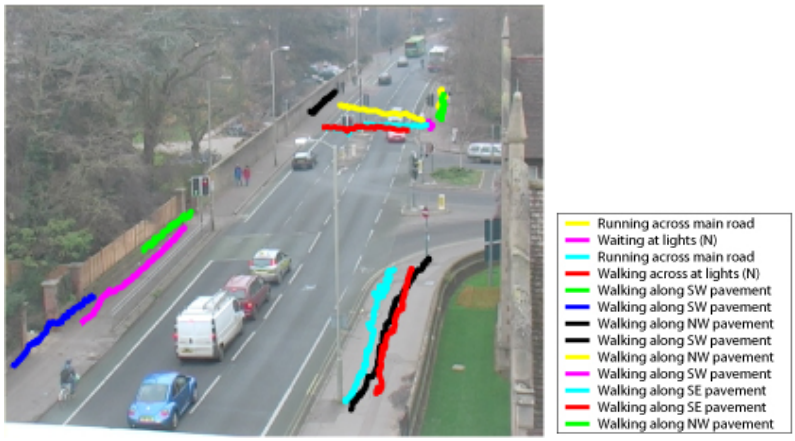


Figure 4.16: (*Left*) A subset of the trajectories in the exemplar data representative of expected activity in this urban scene are shown here. (*Right*) The labels corresponding to each example are shown in the legend.

Figure 4.17: Instantaneous actions recognised in this scene.

Person crossing the road at traffic lights

Frame 380 Frame 437 Frame 468 Frame 646

Frames	Estimated activity
380-427	NE-pavement, walking
428-486	N-ped-crossing, walking
487-511	*N-ped-crossing, stopped*
512-632	N-ped-crossing, walking
633-670	NW-pavement, walking

Figure 4.18: The text commentary for a person crossing the road at a set of traffic lights. From frames 487 to 511 the traffic lights obscure the person (tracking continues because feet are visible) and the motion-type is incorrectly estimated.

Person jogging across road

Frame 5336 Frame 5381 Frame 5462

Frames	Estimated activity
5330-5470	road, running

Figure 4.19: In this example the priors are critical to the choice of the correct spatio-temporal action. Running is not represented as often in the example database. Therefore if the priors for each simple-action are computed on the basis of frequency then the ML spatio-temporal action for this sequence is *road, walking*. If however, the priors are uniform the ML result is as shown here. Note that in either case the correct activity is still represented in the distribution over spatio-temporal actions.

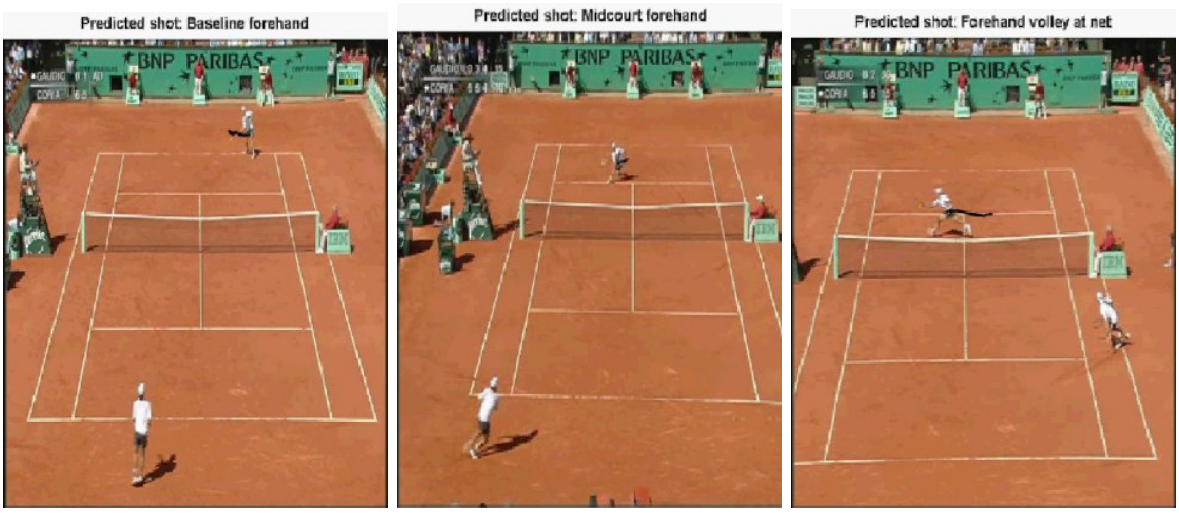


Figure 4.20: Shot matching in tennis sequences.

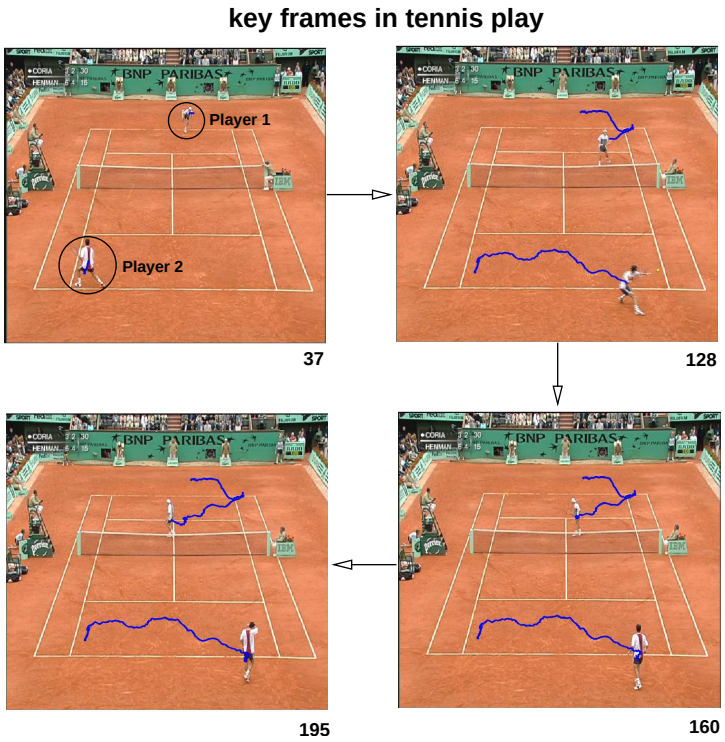
the camera (nearside) which significantly reduces ambiguity in the choice of simple-action (a backhand by a player facing one direction is, motion-wise, very similar to a forehand from the other viewpoint). Taken with the labelled position examples *baseline*, *midcourt*, *backcourt* and *net*, we have 33 possible actions for each player. Testing is performed using previously unseen footage from a 5th match involving two previously unused players. Figure 4.11 shows an example of the spatio-temporal action selection performed by the first two levels of our system. Note that although the figure shows the maximum likelihood estimate, the system does retain a distribution over possible spatio-temporal actions.

Individual shots are matched using the action-recognition method, as shown in Figure 4.20. An entire tennis play "commentary" is generated in Figure 4.21.

Overall detection rates for all of the sequences are shown in Figure 4.22.

4.5 Conclusion

In this chapter a method for action recognition is reported. The particular features we have chosen to use to construct a feature-level description are easy to obtain and photometrically invariant, but one is certainly not limited to these features. The inclusion of a description of local motion raised three issues:



Frame	Estimated shot	Likelihood	Ground truth
32	P2: *Baseline forehand*	0.41	Serve
39	P1: Walking at baseline	0.54	Walking at baseline
51	P1: *Running at baseline*	0.60	Running at midcourt
76	P2: Walking at baseline	0.36	Walking at baseline
99	P2: Walking at midcourt	0.45	Walking at midcourt
132	P2: Backcourt forehand	0.66	Backcourt forehand
148	P1: *Forehand at net*	0.34	attempted forehand

Figure 4.21: A text commentary for selected frames of this tennis play. Where the estimated shot deviates from the ground truth it is marked in italics.

Sequence	% detection, ML model correct	% detection, true model in distribution
Urban street	96.7	100.0
Junction surveillance	74.0	89.5
Tennis	59.4	88.8

Figure 4.22: The detection rates for the three video sequences used in this chapter.

1. Searching a large database effectively,
2. Ensuring temporal consistency of model choice when the example data is sparse,
3. Combining independent descriptions of action in a principled way to describe action and behaviour.

In this chapter we have combined disparate ideas from the literature for each of these problems in a novel way and the results demonstrated the efficacy of these solutions.

We showed that by creating a framework for the propagation of uncertain information in a principled fashion coupled with a method for incorporating expert domain knowledge it is possible to classify human action non-parametrically and deal with ambiguity. Though we have demonstrated the system with application to video annotation, we could equally apply the techniques to abnormality detection. Video annotation and/or novelty detection are simply means to a grander goal of developing a system which can *explain* what is being observed, not simply *detect* what has been previously observed.

In summary, the work presented in this chapter has made the following contributions:

- Recent results in data-driven human action recognition [44] have been extended. We have explicitly shown that a concatenated local motion descriptor gives more effective discrimination in smaller datasets by improving temporal context,
- By representing position and velocity, in addition to local motion, spatial context is given which is important for higher level reasoning,
- Inspired by Sidenbladh's [144] method for generating a set of particles representing a distribution over trajectories, we structure the search over actions using a PCA decomposition of the database. This yields an efficient search which is $O(logN)$ compared with $O(N)$, which for our application means 20x faster than for nearest-neighbour) and additionally by including a stochastic element to the search we can easily obtain a likelihood distribution over possible actions,

- The use of a Bayes net for fusion of non-parametric database search results for action recognition
- Human level descriptions are achieved by abstracting the actions as a precursor.

Chapter 5 addresses the smoothing of action sequences using the basic rules of the scene in order to produce a more robust text commentary of observed activity. We also consider higher level reasoning about scene context by representation of behaviours as action sequences, with representation and recognition of behaviour achieved via HMMs.

5

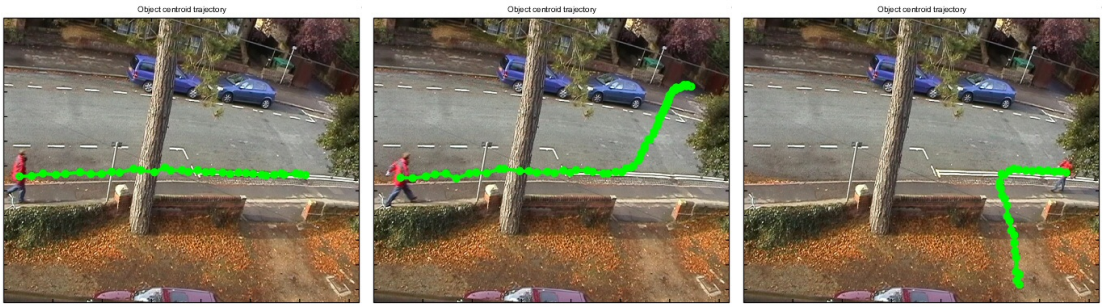
Behaviour recognition

In this chapter we develop a system for human behaviour recognition in video sequences. Human behaviour is modelled as a stochastic sequence of actions. We provide justification for our novel approach by comparing it with motion modelling techniques such as Kalman Filtering. HMMs which encode the rules of the scene are used to smooth sequences of spatio-temporal actions. The inputs to the HMM are actions rather than raw image data such as pixel coordinates. High-level behaviour recognition is achieved by computing the likelihood that a set of (additional) predefined HMM explains the current action sequence. Thus, human actions and behaviour are represented by a hierarchy of abstraction: from person-centred actions, to actions with spatio-temporal context, to action sequences and finally general behaviours. We demonstrate the results on broadcast tennis sequences and urban surveillance footage for automated video annotation.

The work described in this chapter was published in the proceedings of the International Conference on Computer Vision, Beijing, 2005 [131], the proceedings of Imaging for Crime Detection 2006 [134] and has been submitted to the journal Computer Vision and Image Understanding [133].

5.1 Can Kalman Filters model high-level behaviour?

From video of an urban street, three Kalman Filter motion models are learned using EM (see appendix C) from tracked object data. These activities are selected from tracked image data which correspond to the following behaviour and labelled by hand as (a) "walking on the pavement", (b) "crossing the road", (c) "turning left into the driveway". Examples of the training data used to learn these activity models are shown in Figure 5.1.



(a) (b) (c)

Figure 5.1: Input data: the image coordinates from these tracked sequences were used to train three models which we specify as normal for this particular scene. (a) "walking along pavement" , (b) "crossing road" , (c) "turning into drive".

The filtered state estimates and the forward predictions, based on the learned model for each of the sample motions, are shown in Figure 5.2.

The initial settings relating to the standard Kalman Filter equations provided in appendix B section B.1.3 are as follows:

$$
\begin{aligned}
F &= \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \\
Q &= \begin{pmatrix} 0.1 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0.1 \end{pmatrix} \\
H &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \\
R &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}
\end{aligned}
\tag{5.1}
$$

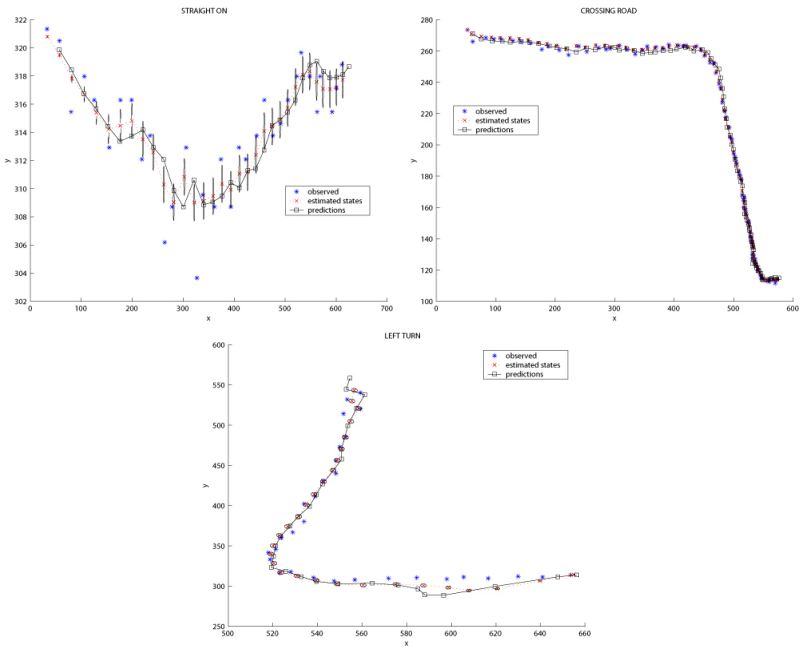


Figure 5.2: Learned motion models, estimated states and predictions for the 3 exemplar models.

F is the process matrix relating the (internal) state at t to the state at $t+1$, Q is the process noise covariance, H is the measurement matrix relating the state to the measurement and R is the measurement noise covariance.

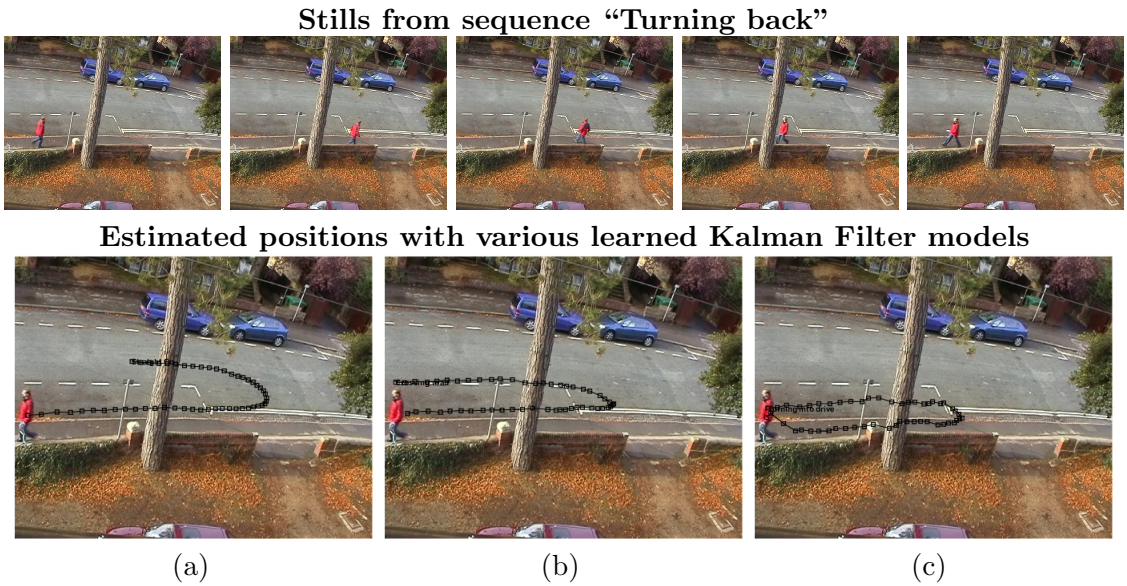
The state, $\mathbf{x}$, contains position and velocity; the measurement, $\mathbf{z}$, measures position only:

$$\begin{aligned} \mathbf{x} &= \begin{pmatrix} x & y & dx & dy \end{pmatrix}^T \\ \mathbf{z} &= \begin{pmatrix} x & y \end{pmatrix}^T \end{aligned} \tag{5.2}$$

The initial values are given by $dx = 1$, $dy = 0$, $[x, y] = \mathbf{z}_1$.

5.1.1 Model selection

The input example in Figure 5.3 is of a person turning sharply back in the direction from which he/she came. From the results shown in Figure 5.3, it would appear that each of the models would track reasonably well until the physical turning-point at which point the person turns sharply. When the motion changes, the models predict differently. It is expected they all would recover, some more quickly than others. Figure 5.3 illustrates the crucial difficulty



(a) (b) (c)

Figure 5.3: This Figure explains why the Kalman Filter approach is not appropriate. (a) The exemplar model "walking along the pavement" is propagated forward in the first figure. The response is quite slow by comparison to the swiftness of the turn. The track suggests we may have reached the minimum of the error and the innovation would increasingly recover beyond the final frame in the sequence. (b) The second exemplar model, "crossing the road" (*centre*) is clearly more effective i.e. the errors are smaller. (c) The third plot (*right*) shows the model "turning into drive" which is most significant because it tracks effectively (certainly with the least error of all in the bank of models).

with the Kalman Filter for parameterising high-level concepts. The discriminative power of the exemplar models is weak when we are interested in model-selection and spatial information.

To test model selection using the Kalman Filter parameterisation we chose another sequence which is unusual (Figure 5.4), compared to the exemplar data, and a sequence which represents normal activity but with some ambiguity (Figure 5.5). The best-fitting model is chosen on the basis of the log likelihood output of the Kalman Filter using a particular model at every time step i.e. a ML estimate (the model order is identical). The log likelihood is not a sum over all the previous data, rather the likelihood score at the last data point[1].

When new tracker data arrives, a set of Kalman filters, one for each of the models in the training set, is used to determine the most likely model at that time step.

The log-likelihood of a model explaining the data is calculated using all the data at the given

[1] Although giving the tracker some "memory" by summing the likelihoods over all time steps may be useful for model selection with regard to the entire sequence and not simply an individual time step.

time step (e.g. *top left* graph in Figures 5.4 and 5.5). The likelihood of one update step is

$$\Lambda = \mathcal{N}(e; 0, S) \tag{5.3}$$

where S is the covariance of the innovation (e) denoted by:

$$S = HVH' + R \tag{5.4}$$

with:

$$V = var(\mathbf{x}_t|\mathbf{y}_{1:t}). \tag{5.5}$$

Now the total log-likelihood is calculated as the sum at each time step as $\sum_{k=1}^{T} \log \Lambda_k$. This value is plotted in the *top-right* graphs in Figure 5.4 and 5.5. This value, therefore, contains a memory of all time steps up to that point. The cumulative log-likelihood will always decrease since the log-likelihood at any time-step is negative but it nevertheless provides an indication of how likely it is that the model in question produced the entire dataset.

The innovation of the Kalman Filter can be viewed as a measure of the tracking error i.e. the difference between prediction and observation. This is plotted in the *bottom-left* graph of Figure 5.4 and 5.5. This is the most useful graph to study because the innovation will increase if the data is not explained well by the model. It is expected however that the tracking error will reach some minimum value and then recover. The graphs in Figure 5.4 and 5.5, and of the log-likelihood of the ML model in the exemplar set producing the data for each frame (*top-right*) and the log-likelihood for every model in the exemplar set (*bottom-right*) indicates how well the prediction is performing at each time step.

Model selection in the case of a novel activity

One definition of "novelty" in the context of video is that which is not represented by the "normal" training data. A novel example, in the context of the data which is the subject of this

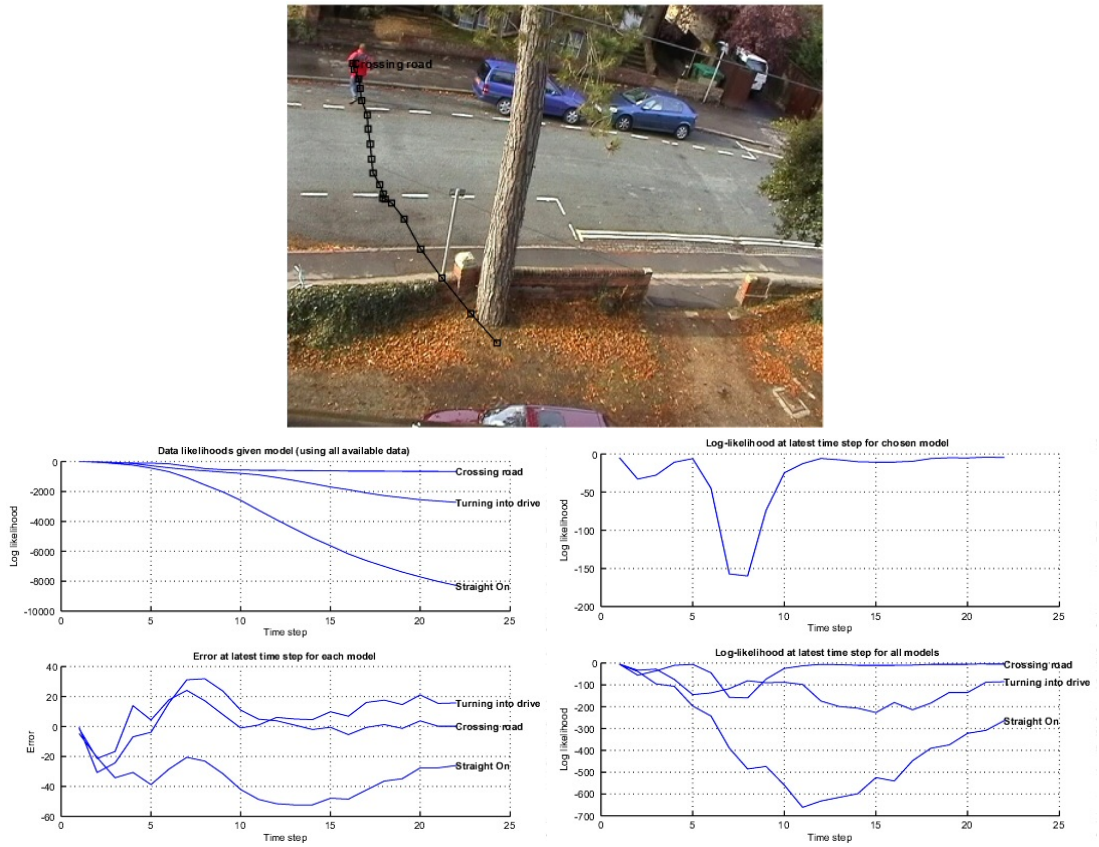


Figure 5.4: Model selection in the presence of a novel input: (*clockwise from top-left)* 1. The plot of log-likelihood given all the data available at each time step is a good indication of how well the data fits the model overall; 2. For the chosen model (one with the highest likelihood at that time step) the likelihood is plotted. There is a clear anomaly at around frame 7; 3. The log-likelihood at each time would be expected to be correlated to the plot of tracking error at each stage (*bottom-left*); 4. (*bottom-right*) The log-likelihood at each stage for each independent model is shown.

set of experiments, corresponds to the activity "running across the road" (Figure 5.4). When this input is encountered, the "crossing the road" exemplar model has the highest likelihood of explaining the data after frame 10 as is shown in Figure 5.4. Despite the similarity of the label "crossing the road" to the observed action, they are actually quite different. (The "crossing the road" action in the exemplar set is shown in Figure 5.1 and can be contrasted with Figure 5.4.) Moreover, it is where the activity occurs that makes this example "interesting". However, model selection of this type does not allow us to effectively discriminate between models which are incorrect spatially or incorrect due to the incremental motion (e.g. constant velocity vs. constant turn).

Model selection in the presence of ambiguity

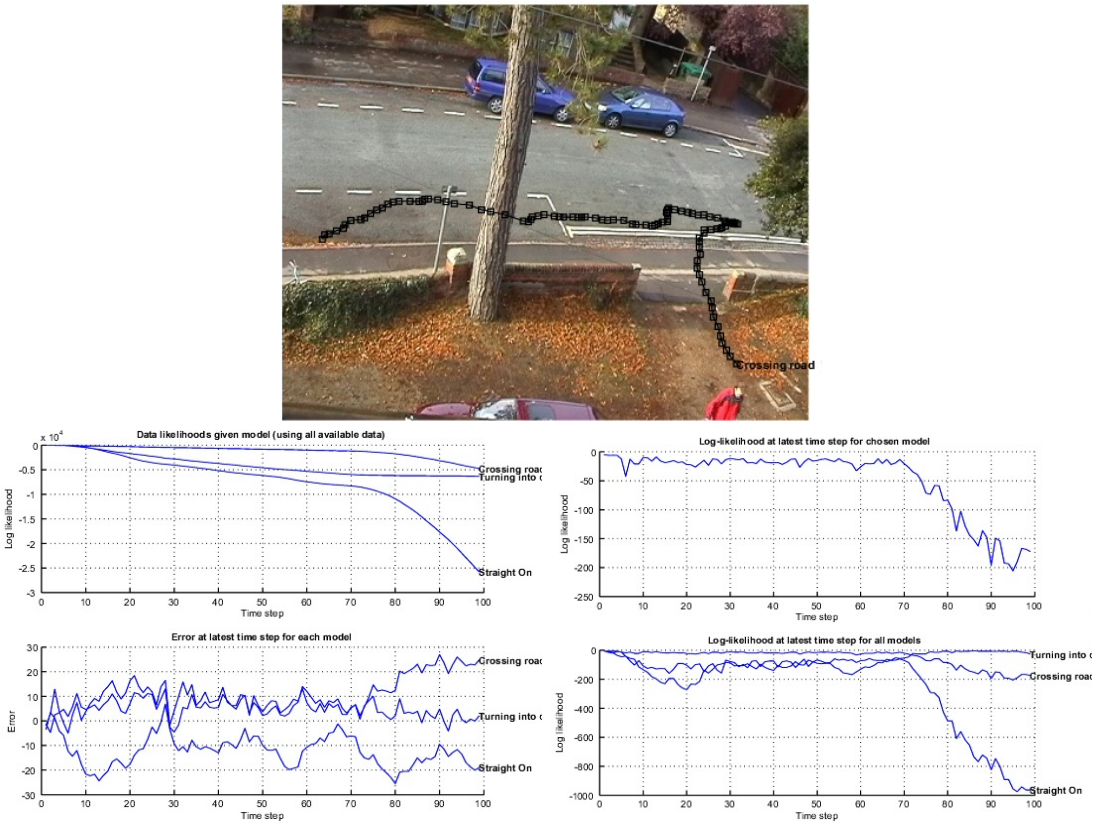


Figure 5.5: Model selection in the presence of ambiguity. There is a low likelihood even for the models from the bank which, to an expert, are most appropriate for tracking this data. Moreover despite the likelihood values improving (i.e. the innovation decreasing) in the model "turning into drive" it is clear this is not the correct description of the activity. The total log-likelihood reflects this mismatch.

The input in Figure 5.5 is similar to the exemplar model "walking on the pavement", shown in Figure 5.1, in terms of spatial position, but quite different in that the velocity is staggered and erratic. This behaviour would have the human description "dithering" or perhaps "wandering". A human would have no trouble spotting this as somewhat abnormal, if not suspicious. In Figure 5.5, the plot of log-likelihood of the data fitting each model (*bottom-right* graph) shows there is ambiguity about which model is appropriate at certain points in time. As can also be seen in Figure 5.5, no model is appropriate for the entire length of the sequence but there are places where one model is a better explanation of the data than the others, particularly after frame 70, which corresponds to turning into the drive, which is a right turn.

This result suggests that more complex actions may be identified by sequences of motion models

which are simpler than those which represent complex behaviour. That is, tracking at one level of abstraction higher than the image data but not at the level of extended behaviour comprised of multiple actions. For example a left turn could comprise a constant velocity model followed by a constant turn model, which essentially requires that the model allows for changes in state.

5.1.2 Lessons

Without wishing to over-state the importance of this experiment, it does highlight an interesting point which must be considered when using state-space models, and provides some additional motivation for the approach we take in this chapter.

The Kalman Filter does not capture the distinctions required for labelling video such that reasoning about scene activity could be achieved. The point of interest in the analysis of the trajectory is where the likelihood of the data being described by the model decreases rapidly. This point is most clearly reflected, in this case, by the innovation of a Kalman Filter. The essential difficulty is that the error recovers as the motion estimate improves over time, as we showed in Figure 5.3. The resultant ambiguity in model selection - do we choose the model performing best at a given time step, or overall? - means that it is not possible to reliably differentiate between a normal and an unusual activity, even in this simple case where the activities are quite different.

The exemplar behaviours for which we learned model parameters are composed of sequences of simpler actions. But the Kalman Filter has one state, the value of which varies over time, according to a predefined distribution. The HMM, by contrast, allows state changes but requires large quantities of good quality training data for each atomic action, i.e. data where the state transitions we want to learn are well-represented, as we have discussed in the literature review of chapter 2.

Therefore, we propose a new technique for incorporating higher-level knowledge about behaviour is required, which we now introduce.

5.2 Behaviour as a sequence of actions

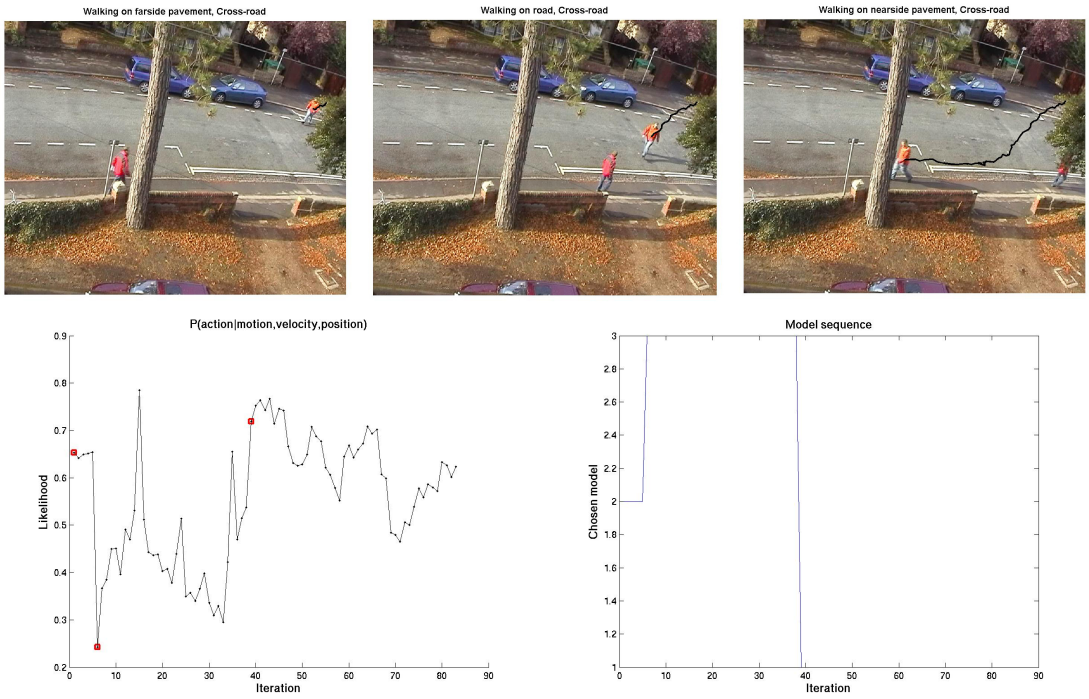
The action-recognition work of chapter 4 abstracted image-level information into a discrete distribution of intermediate-level actions with high-level labels, of the type an expert analyst would be expected to provide. Taking the ML action from this distribution over all possible actions at each frame, this constitutes a ML text-commentary on video. This is of interest in its own right, and has a number of applications.

We now go one step further. By utilising the knowledge of the "rules" of a scene and encoding those rules as a HMM, where the inputs to the HMM (indeed the hidden states) are indices into specific spatio-temporal actions with the associated likelihood. This has three immediate benefits. First, the prior knowledge of a human analyst can be quickly and correctly incorporated into a probabilistic reasoning system. Second, there is no need for large quantities of training data to provide us with the models of behaviour. Third, even if global behaviour is inaccurately estimated using this technique, the intermediate (spatio-temporal action) level, still provides a good description of activity, as we have seen in chapter 4.

5.2.1 Behaviour parameterisation

In a scene that is well-understood, for example, an urban environment which features traffic and pedestrians, the global behaviour "crossing the road" would be an activity one may wish to detect. It is clear that, given the indices of the spatio-temporal action, one could write down the expected action sequence which would constitute this behaviour. A parse of such behaviour is shown in Figure 5.6. This action sequence can therefore be encoded in a Markov Chain by writing down the expected transition matrix. The states of such a transition matrix have the advantage of being direct representations of the (indices) into spatio-temporal actions plus likelihoods.

It is clearly a straightforward matter to specify the HMM parameters for a specific behaviour when the parse of that behaviour, in terms of the spatio-temporal actions, is known. HMMs for a set of normal behaviours in this scene can be defined. In total then, for this scene, we



Parse of behaviour into spatio-temporal actions	
Frame	**Action**
1-25	Walk-farside-pavement
26-195	Walk-road
196-350	Walk-nearside-pavement

Figure 5.6: The graph at the bottom-left shows the likelihood of the ML spatio-temporal action at each frame for the tracked person (top row), change-points are shown in red. Beside this the ML model sequence is plotted, representing an automatic parse of the global activity (*crossing-the-road*) into its constituent actions.

defined the following behaviour HMMS:

1. **Walking-along-nearside-pavement.** The parse of this behaviour is a continuous sequence of the spatio-temporal action, walk-on-nearside-pavement.

2. **Turn-into-drive.** The parse of this behaviour is the sequence: walk-nearside-pavement → walk-in-drive.

3. **Cross-road.** This is composed of the following action sequence: walk-farside-pavement → walk-on-road → walk-nearside-pavement.

The inputs to the HMM are two-vectors containing the index into the spatio-temporal action, and an associated probability of that action. The observation probabilities are discrete and the output of each state is the index into a spatio-temporal action (with associated likelihood). So, for example, for the behaviour "Cross-road", above the parameters of the behaviour HMM are specified as follows:

$$\begin{aligned} \Pi &= \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \\ A &= \begin{pmatrix} 1 & 0 & 0.4 & 0 \\ 0 & 0.6 & 0 & 1 \\ 0 & 0.4 & 0.6 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \\ B &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \\ Y &= \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \end{aligned} \tag{5.6}$$

Where Π is the matrix of priors, A is the state transition matrix, B is the observation matrix, and Y is the outputs from each state. The states in this example correspond to:

1. Walk on the near-side pavement,

2. Walk on the far-side pavement,
3. Walk on the road,
4. Walk in the driveway.

In the above example, the interpretation of the state transition matrix, A is:

- When walking on the near-side pavement (state 1), the person will stay on the near-side pavement,
- When walking on the far-side pavement, the person will most likely to stay walking on the far-side pavement (state 2), but a transition to the road (state 3) is allowed,
- When walking on the road, the person will most likely stay walking on the road (state 3), but can move to the action walking on the nearside pavement (state 1)
- When the person is walking in the drive (state 4), no transitions are allowed as this action is not expected to occur.

Similarly, behaviour HMMs are specified for the other behaviours, "Walking-along-nearside-pavement" (which is quite trivial, being a continuous sequence of walking-on-pavement actions) and "Turn-into-drive".

The advantages of using HMMs are clear. First, they allow us to maintain the probabilistic analysis we have achieved through the action-recognition techniques (even though we use the ML spatio-temporal action). Second, there is a well-understood set of tools for DBNs and HMMs in particular, as discussed in chapter 2. Third, they naturally encode rules, which are vitally important for the automatic recognition of human behaviour extended over time.

The ML sequence of actions and their likelihoods over a number of time steps is used to find the most likely behaviour by computing the likelihoods of each of the predefined normal-behaviour HMMs explaining the current action sequence. Since more complex models generally explain data better we use a likelihood ratio to compare competing behaviour models. The likelihood

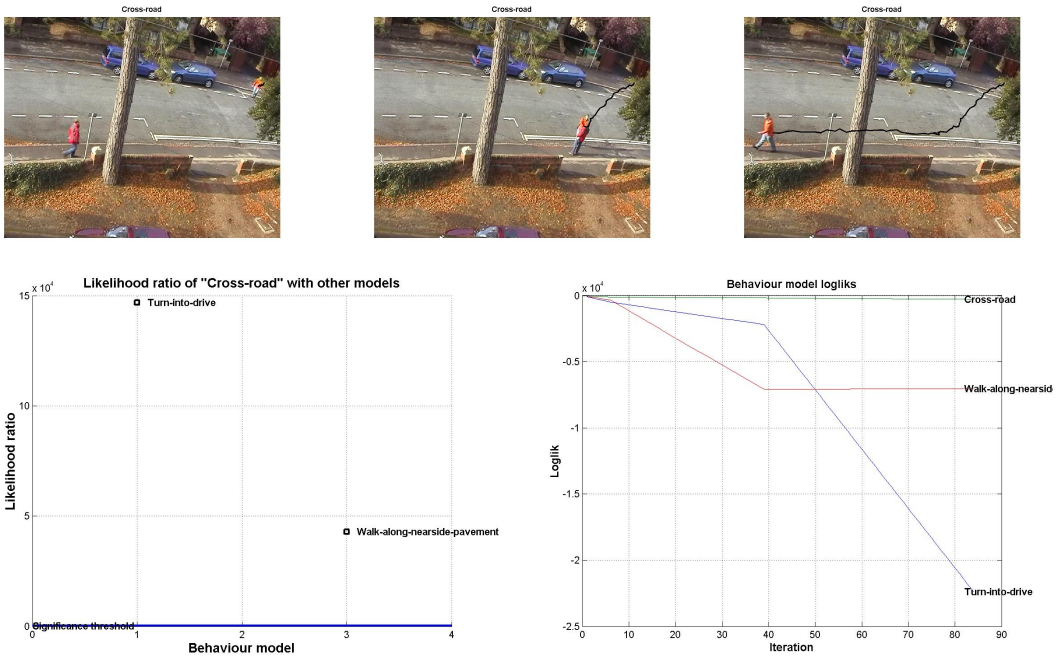


Figure 5.7: The first row of images shows the key frames of an input action, automatically estimated for the tracked person. The corresponding likelihood ratio of the most likely model with the other behaviour models in the bank of models is shown in the second row (*left*). In this case, the behaviour is correctly classified as *cross-road*. The final graph shows the likelihood of the behaviour model HMMs over the entire sequence (*right*).

ratio for comparing two hypotheses H and H' with probabilities $p(H)$ and $p(H')$, respectively, is computed as:

$$LR = 2(\log(p(H)) - \log(p(H'))) \tag{5.7}$$

which has a chi-squared distribution parameterised by the difference in the model order. If LR is greater than the 95% confidence value of the chi-squared distribution for $\delta = |O(H) - O(H')|$, the result is statistically significant[2].

An example of high-level classification of a new input activity is shown in Figure 5.7.

[2]If the result is *not* statistically significant then the natural interpretation is that there exists some other, unknown model which better explains the data.

5.2.2 Novelty detection

We propose that there is no need to model unusual activity when normal activity is known, as we show in Figure 5.8. In this case, a set of trajectories are automatically collected and those which correspond to normal behaviour are labelled accordingly. When the latest activity is observed, by comparing the Euclidean distance between the observed trajectory data and the normal exemplar data it becomes clear when an abnormality is being observed, provided the anomaly is significantly different from normal. Using the same principle, the behaviour models we have specified correspond to only normal, or expected behaviour for this urban scene. By setting a lower-bound on the log-likelihood that a given HMM explains the current action-sequence it is therefore possible to detect anomalous behaviour, as we show in Figure 5.9. Note that, using the Kalman Filter parameters, we were also able to detect an abnormality (as shown in Figure 5.5). However, using the behaviour HMMs we now know *where* the activity takes place and can infer what rules have been infringed. This means it may now be possible to reason about *why* the behaviour is unusual.

5.2.3 The encoding of general scene rules

An advantage of this method is that it is considerably more general than learning examples of global behaviour direct from trajectory data. This is because the action-recognition stage, through computation of the distribution over the raw training data, abstracts us from the training data itself, allowing the behaviour HMM to be a general representation of scene rules at a high-level. This is illustrated in Figure 5.10 where two examples of the same behaviour are enacted. They are both classified correctly, as anyone with knowledge of this scene would confirm, but the low-level data itself is very different, as can be seen from the trajectories in the image plane. If one learned models directly from the trajectory data not only would considerably more training data be required but two HMMs would have to be learned (constituting a significant increase in training data in itself), as opposed to one general HMM, which can be readily specified, with our approach.

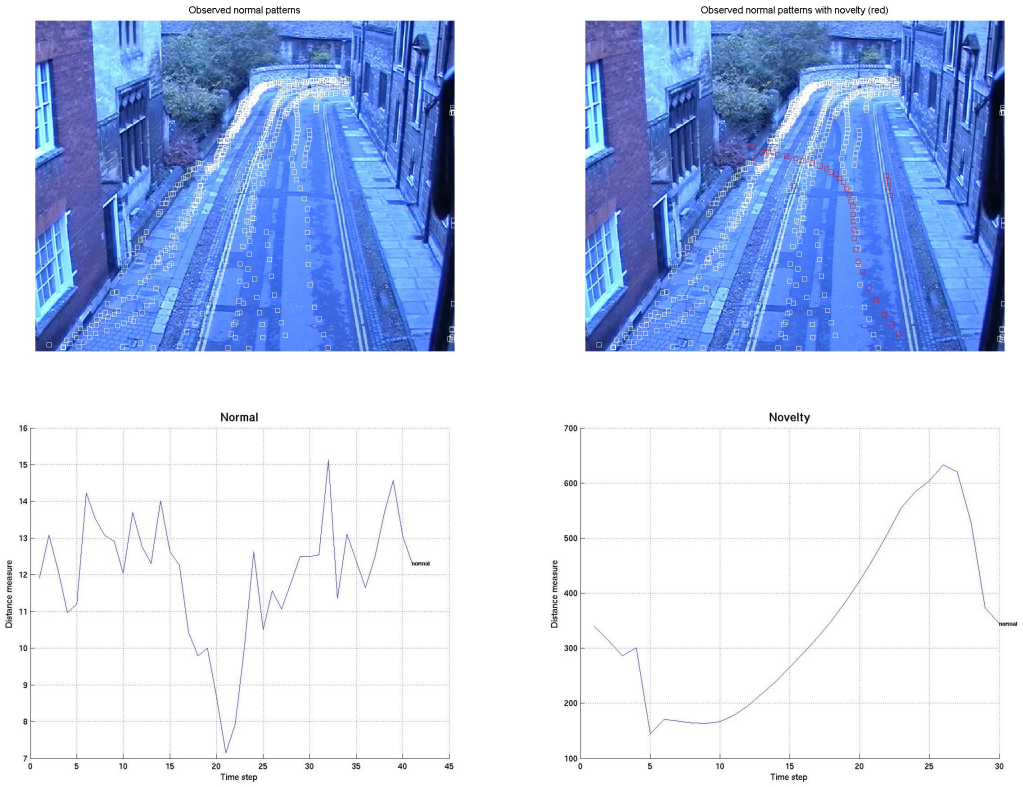


Figure 5.8: Detecting abnormality by modelling only normal activity. Here, a simple sum of squared distances metric of the input to the training data (coordinates) is computed for a normal input (*bottom-left*) and an anomaly (*bottom-right*).

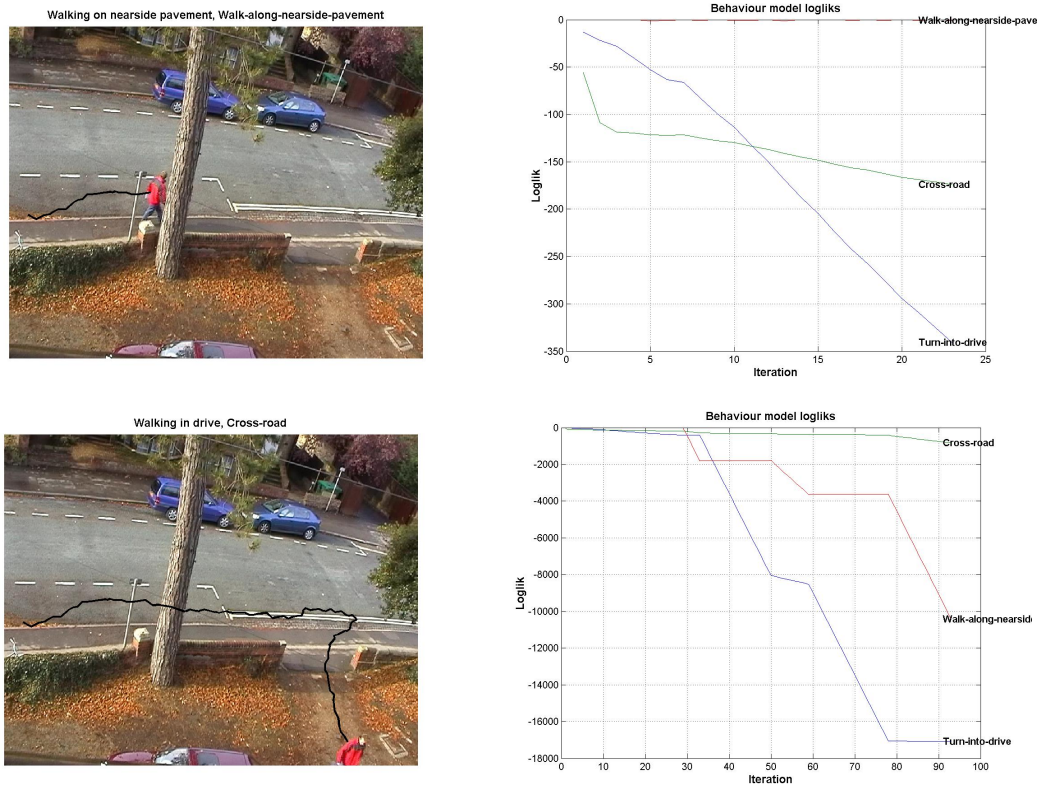


Figure 5.9: This is an unusual action which shows how novelty detection is possible using our method. Near the start of the sequence (*top-left*), the behaviour is correctly identified as *walking-on-nearside-pavement* (*top-right*). As the activity evolves, it becomes apparent that the behaviour is somewhat unusual and this is quite clearly reflected in the likelihood of the normal behaviour models explaining that sequence (*bottom right*). The likelihood scores become so low that the only explanation is that there is another, abnormal model which explains the data.

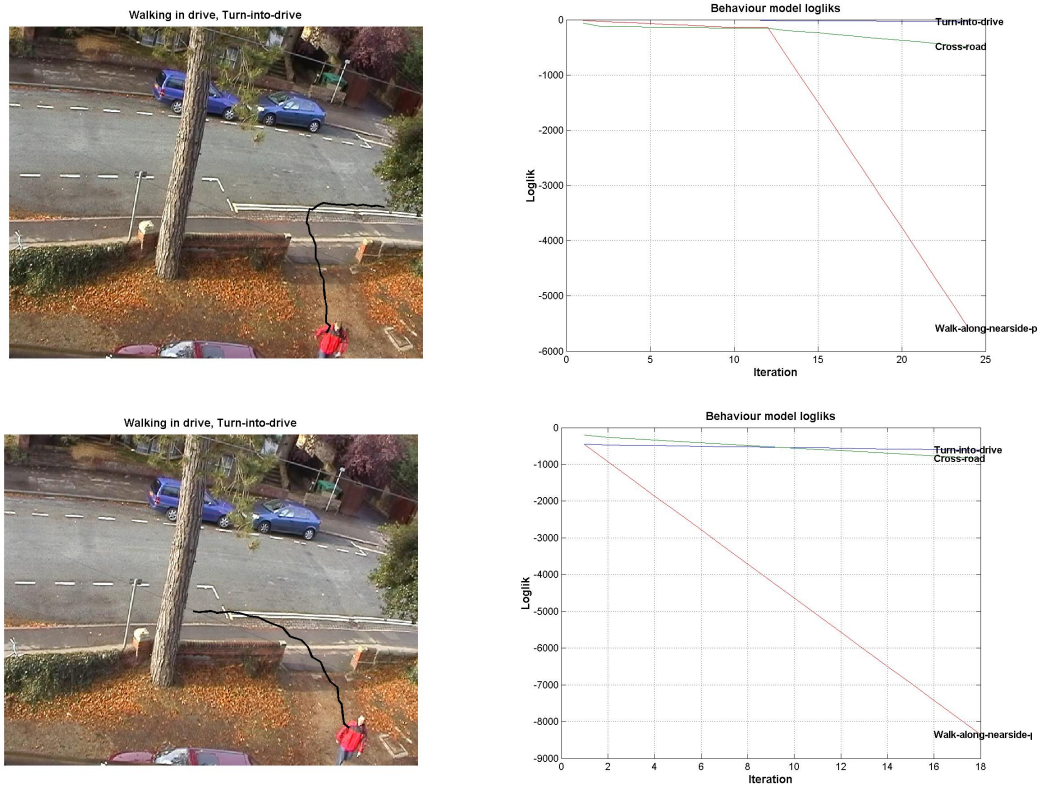
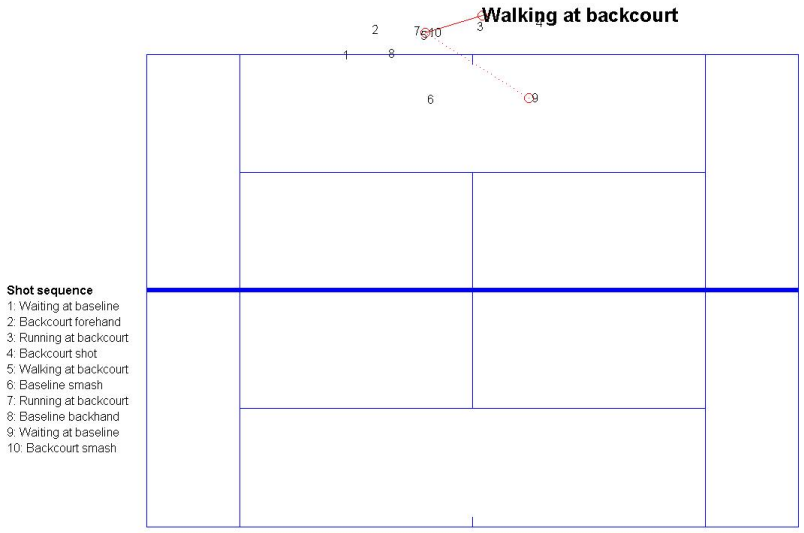


Figure 5.10: The generality of the behaviour recognition method is demonstrated in this example. Here, we show the detection of the behaviour exhibited when someone is observed to walk down the pavement and turn into the driveway of the house. The top-left example is parsed into its action sequence and a HMM is specified from this model sequence. The likelihood of each behaviour HMM is shown beside the sequence. The HMM associated with the *turning-into-drive* behaviour is used to classify the same behaviour but performed in different ways, as shown on the bottom row.

5.3 Improving tennis commentary using known player-types



Serve-volley player	Baseliner player
Service	Waiting at baseline
Waiting at backcourt	Backcourt forehand
Backcourt backhand	Running at backcourt
Walking at baseline	Backcourt shot
Midcourt Backhand	Walking at backcourt
Running at net	Baseline smash
Net volley	Running at backcourt
Running at net	Baseline backhand
Net volley	Waiting at baseline
Running at net	Backcourt smash
Net volley	[END POINT]

Figure 5.11: A simulated play between a baseliner player and a serve-and-volley player (*left*) using the respective HMM tennis player "agents" is shown in the table. The top picture shows the commentary (*left*) with generated positions of the baseline player superimposed on a court.

As detailed in chapter 4, spatio-temporal action sequences are computed for each player in a tennis match. This action sequence is used as the basis for deciding what type of global behaviour is occurring for each player. The behaviour types, which we specify in advance by encoding the expected transitions in a HMM, are *baseline-rally* and *serve-and-volley*.

A text commentary is obtained from the first two levels of our system by simply selecting the ML action at each instant. This however neglects that in many scenarios domain knowledge can be used to improve these estimates. For example, a *service* shot could easily be confused

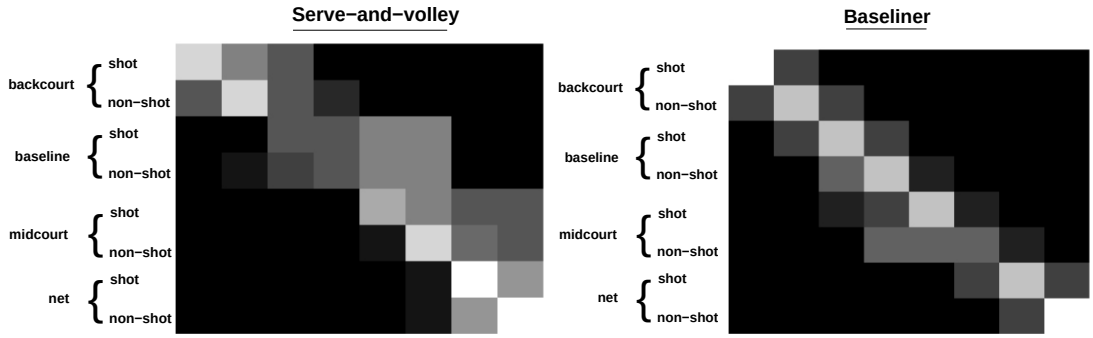


Figure 5.12: The basic rules which dictate likely transitions between court positions for the serve-and-volley (*left*) and the baseliner (*right*) player are shown here. The serve-volley HMM encodes a preference for playing at the net, while the baseliner prefers to stay at the baseline, but will stay at the net if forced there by the opponent.

with a *baseline-smash* if it were not known that a *service* only occurs at the start of a point.

Since the series of expected shot *types* is well-established we smooth the shot commentary using a HMM which encodes some specific rules. These rules are:

- A service starts a point,
- The player who is not serving waits at the baseline,
- A shot exists for a typical number of frames,
- position on the court must go through physically possible transitions (e.g. midcourt is *en route* to the net from the baseline),
- A serve-and-volley player tries to move to the net and there is only a small probability of returning to baseline when he/she has advanced,
- The baseliner player prefers to return to baseline if forced to midcourt but if at net will prefer to stay in the advanced position,
- Each player is expected to make regular transitions between shots (e.g. service) and non-shots (e.g. running).

The position rules are encoded in the HMM state-transition matrices which are shown in Figure 5.12, as is the transition between shots and non-shots at these positions. The observation

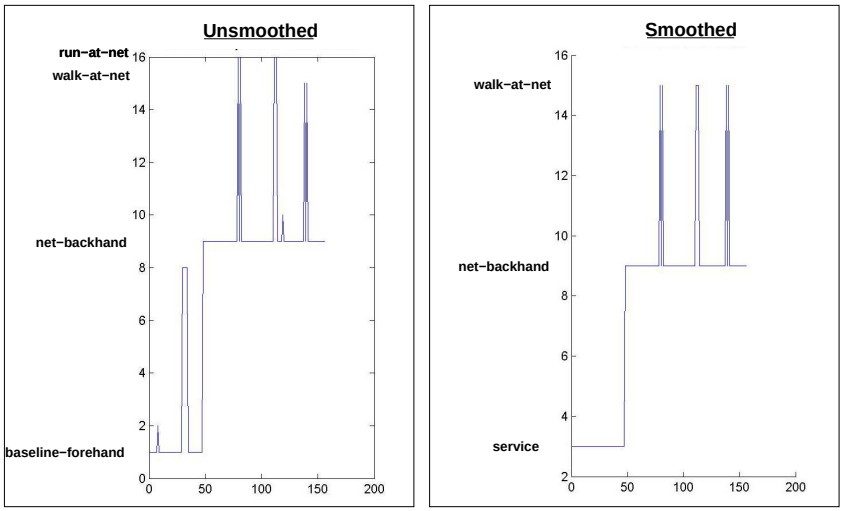


Figure 5.13: Smoothing the shot sequence which arises from the spatio-temporal action-recognition phase (see Figure 4.1) provides consistency across the shot choice and allows important expert knowledge to refine the shot selection. In this example here the player is known to be serving and HMM for a serving player is used to smooth the shot sequence. The improvements can be seen by comparing the unsmoothed (*left*) and smoothed (*right*) sequences in particular the serve is no longer omitted and the shot to non-shot transition is observed.

probabilities are uniform distributions over shot-types, except in the case of a serve-volley player where, initially, the positioning at the baseline indicates a service is the current shot. Therefore, given a smoothed, MAP position estimate using the HMMs, and a choice of shot or non-shot, smoothed spatio-temporal actions can be generated. A simulated tennis play, generated using these rules, is shown in Figure 5.11.

Results of shot-matching and the resulting ML commentary are shown in Figure 5.13. As can be seen from the smoothed shot sequence in the graphs of Figure 5.13, the improvement can only come from: (a) the smoothed positions; (b) the supervision of shot vs. non-shot by dividing the indices into the spatio-temporal actions into shot/non-shot groups. The best estimate of each shot or non-shot action arising from the action-recognition stage is still required, except in where the first shot is constrained to be a service.

Analysing data from 4 tennis matches, and using these known player types, we find this expert knowledge yields a considerable improvement as the results in the table in Figure 5.14 indicate.

ML action (% correct)	MAP smoothing (% correct)
59.4	88.8

Figure 5.14: The effect of smoothing the ML sequence using an HMM which encodes expert knowledge about tennis matches.

5.4 Conclusion

In this chapter we have demonstrated that behaviour can be modelled as a stochastic sequence of actions. This observation has allowed us efficiently to encode expert knowledge about the scene in a HMM by writing down the expected state transitions and associated probabilities.

The HMM representation of behaviour subsequently enables classification of normal activity in a probabilistic and principled fashion. By modelling only normal activity, abnormalities can be detected by the fact that none of the set of normal models explains the observed action sequence well.

The use of known rules to generate smoothing HMMs for more complex situations, specifically tennis matches, significantly improves the results of the ML action-recognition phase. The complexity of the expert knowledge encoded in the tennis player-type HMMs could be extended to include specific shots as opposed to only positions that we have used here.

6

Causal reasoning

In this chapter we draw the results of the preceding chapters together to achieve the goal of this thesis: causal reasoning about human activity in video. In particular, we discuss an agent representation for modelling individual human behaviour observed in surveillance video. The activity estimates obtained from the work of chapters 3, 4 and 5 are used as the information available to the sensors of the agent. Further, to demonstrate the utility of these intermediate descriptions of activity we show how a set of rules, articulated at a human-readable level, achieve causal reasoning about the activity of multiple agents. By specifying a reasoning process which is initiated by certain events, a general technique for causal reasoning in video is demonstrated. We show results in two scenarios: tennis and urban surveillance. Although, in this chapter we predominantly describe a reasoning process which is Maximum Likelihood, we discuss, with an initial example, how this could be extended in future work to fully Bayesian reasoning.

The work of this chapter has been published in the proceedings of Imaging for Crime Detection and Prevention 2006 [134].

6.1 Introduction

In a multi-agent environment, an agent's actions can be modelled as a consequence of what they sense and what they reason about other agents. As we highlighted in chapter 2, in the work of Dee and Hogg [39] a model of human behaviour is proposed based on the assumption that people move through an urban scene directly towards predefined goals. Then, by comparing how "interesting" the model says the observed behaviour is to how worthy of further investigation a human analyst believes the behaviour to be, the model is verified. Dee and Hogg's work principally relies on inferring what an agent can sense through the projection of rays, based on the centroid trajectory of an agent, and the subsequent use of the goal-directed model of behaviour to predict how the agent is expected to act.

Our goal is, similarly, to confer upon the system the ability to "explain" what is observed and to recognise when the observed behaviour is not explicable. In this work, the reasoning can be performed on the basis of: (i) what is seen i.e. what actions and behaviours are estimated on the basis of the training data; (ii) what beliefs, desires and intentions it is assumed our person-agents have. The latter is normally predefined using an expert's prior knowledge and, clearly, this must vary depending on the context.

We now have a richer set of estimates about a human agent's activity in video, including gaze-direction (as opposed to simply overall heading inferred from body-direction which is the estimate Dee and Hogg obtain), spatio-temporal actions and behaviours extended over time. The low-level vision tools we have developed, therefore, provide many *whats*, i.e. a set of facts about an agent, from which we require to derive one or more *why's* i.e. an explanation of an agent's behaviour.

Before turning to our solution to this problem, we review the scientific state-of-the-art in the area of reasoning about activity in video.

6.2 Review of relevant literature

6.2.1 Causal reasoning

Making sense of a scene can be thought of as:

> Assessing its potential for action, whether instigated by the agent or set in motion by forces already present in the world [18].

In other words, a causal interpretation is most easily and most commonly judged by the motion effects that take place, as we have seen in section 2.1.

There is a history in scene understanding research of analysing static scenes. In the work of Cooper *et al.* [33], for example, the causal explanation of a static scene is found in the answer to the question, *Why doesn't this object fall down?* `MugShot` [33] which can successfully pick up cups filled with hot fluid, is one example of a system which successfully analyses a certain kind of scene in which causal relationships can be learned. This is an example of an explanation-mediated vision system which is well suited to a variety of kinds of perceptual and concept learning and has two important aspects for learning: expectations and explanations. The former, if they fail, are opportunities to learn; the latter provides the context and material for learning. This shows the essential limitations of such a quantitative system: where knowledge runs out the system cannot make sense of the scene and a rule or fix has to be implemented to prevent repeated failure. It is inevitable that certain types of scene will be understood while others will confound the system to the point that it cannot learn.

This concept of explanation-mediated vision differs from the model-based approach also found in the literature [86, 162]. A model is a small, finite description of an infinitely complex reality and is constructed for the purpose of answering a particular question. So, for example, if the question concerns the trajectory of a projectile, the model may describe the object in terms of mass and velocity but possibly ignore air resistance effects if the projectile is only travelling a short distance. The process of using models to reason about scenes is characterised by two major sub-problems, both of which must be solved [86]:

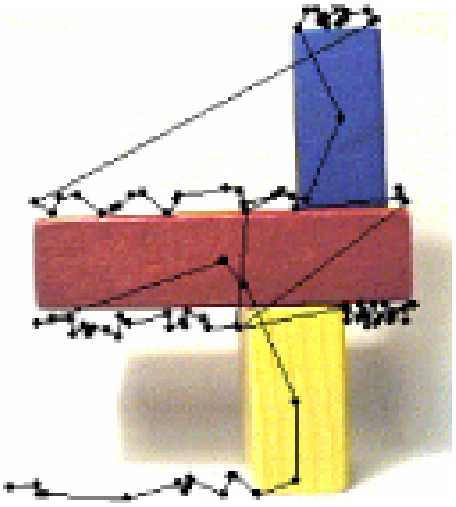


Figure 6.1: From Brand and Cooper [18]. This simple scene is unchallenging as far as extracting image features is concerned. Brand and Cooper demonstrated causal analysis of a series of static images. When the current best explanation fails, more rules are added until a satisfactory automatic "understanding" of the structure (i.e. why the blocks are maintained in static equilibrium) is achieved. Note that the dots shown superimposed on the image are not extracted features but points of visual attention.

1. Selection of the appropriate model or combination of models to answer a given question,

2. Simulation of the model to gain some facts about the world.

The principles of qualitative physics can be extended to encompass motion by priming a system with sufficient elementary causal rules. This is essentially the contribution made in this chapter. However, the literature to be found in the area of qualitative reasoning predominantly deals with analysis of objects in their static state and the predicted motion, or lack of, must therefore be explained in terms of static properties [18]. It should also be noted that the systems Brand and Cooper created were, in addition to understanding static, as opposed to the dynamic scenes we are interested in, well-suited to simple visual analysis. That is, the features required to reason are simple to extract, being centroids and edges of blocks pictured against a clean background, as shown in Figure 6.1.

There is some divergence in thought about how best to represent the knowledge required to develop a reasoning system. By extension from our own experiences, it seems most likely that humans use a combination of model-like simplifications and explanation-mediated learning. Therefore, the approach of learning about a system and then, by extension to many and varied

systems, learning about the world, is one which is naturally appealing. The central question, however, is: *Is it best to model the situation using prior knowledge or should the system be enabled to gather its own knowledge, learning as it goes?* Given that this thesis has, so far, argued in favour of the approach of utilising any prior knowledge at our disposal, particularly in the training phase, we continue in this vein now that we consider higher-level reasoning[1].

The most interesting point to arise from the literature is that one major shortfall in the reported work in this area is the *lack of robust computer vision methods for obtaining low-level information about the agent.* This is a criticism Rigolli who, with Brady, developed a traffic surveillance commentator, identifies [125, 126]. Addressing the issue of obtaining descriptions of agent behaviour is one area with which this thesis has been concerned. We now turn, in this chapter, to demonstrating the efficacy of our previous results by showing how the modelling of human activity is considerably simplified by having an intermediate representation of an individual's behaviour arising directly from the video data.

Let us first consider some specific architectures which are common in the literature for representing and reasoning about this information.

6.2.2 Agents

According to Russell and Norvig, an agent is:

> Anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors [136].

An agent therefore has a series of inputs and a set of actions that can be performed. Consequently, it can be constructed as a software function. When these agents are combined, complex behaviour can emerge which models real-world human behaviour. The work of Andrade and Fisher is a particularly interesting example of this in a surveillance context [1].

[1] That is not to say, that the question of knowledge representation is not an interesting topic for research in its own right, simply that it is beyond the scope of this particular piece of work to discuss.

The complexity of an agent is determined by the kind of environment in which it is found. Russell and Norvig [136] define the following environment properties for an agent:

- **Accessible vs. inaccessible**: Can the agent get complete, up-to-date information about the environment? If so, the environment is defined as accessible.
- **Deterministic vs. non-deterministic**: An agent is deterministic if the next state is completely defined by the current state.
- **Episodic vs. non-episodic**: Can the agent's experience be divided up into unrelated chunks? If it can, it is episodic.
- **Static vs. dynamic**: In a dynamic environment processes other than the agent itself are in operation.
- **Discrete vs. continuous**: A discrete environment has a fixed number of actions that the agent can perform.

There are many types of agent defined in the Artificial Intelligence literature. The most appealing, especially from the point of view of encoding prior knowledge, is the Belief-Desire-Intention (BDI) agent (originally developed by Bratman [23]). This agent has a set of beliefs about its environment. "Desires" are computed on the basis of its goals which, subsequently, dictate its behaviour. Beliefs are theoretical, desires are potential influencers of action and intentions are practical. Since the person agent cannot influence the world, a model of the agent in video must be based on encoding intentions. This type of agent is believed to model decision-making process humans use in every day life [55]. One way to incorporate this kind of agent within a reasoning process is to use a cost function which is dynamically updated on the basis of environment information [124] i.e. to penalise certain types of activity.

Note that the world is only partially observable for an agent. For example, a vehicle driver cannot see some cars due to limitations of view or certain weather conditions. Other drivers' intentions are invisible to him and he has only a stochastic model of the results of own actions.

In order to make decisions, sensor data and a joint probability over sensor data and all possible states of the world is required.

6.2.3 Rule-based reasoning

In *Fuzzy Expert Systems*, Siler and Buckley observe:

> As an expert in your domain, you have probably not found it necessary to formalize your thinking processes, except when trying to explain to a junior person how you reached some conclusion. But the computer requires defining your thinking in some formal way. Various formalisms have been tried. The one which has shown the greatest flexibility and similarity to human thought processes is the rule, although other formalisms have been used, mostly in very special cases. The formalism used by expert production systems is a set of rules of the type:
>
> IF (certain specified patterns occur in the data) THEN (take the appropriate actions, including modifying old data or asserting new data) [146].

Reasoning about courses of action naturally follow from the cost function idea for one agent i.e. the best course of action is followed on the basis of the least costly result. However, agents' behaviour can be reasoned about in a straightforward way using rules which is a more intuitive way to formalise the human reasoning process. Moreover, these rules can be quickly identified and written down by an expert. The following are identified as the positive and negative aspects of taking a rule-based approach [124]:

Pros:

- It is easy to update the system's knowledge by adding new rules without changing the reasoning engine,
- It is easy to transfer between applications by specifying a new set of rules,
- One may embed a large component of domain-specific knowledge,

- Knowledge is contained within an identifiable part of the system,
- An explicit representation of the decision-making steps means that the inference sequence can be explained to the user,
- The computation is data-driven i.e. new data drives the action selection process, thus appearing intelligent.

Cons:

- The complexity of the system can become quite high even for simple actions,
- It is not easy to reason under uncertainty as "hard" decisions are being made.

The conclusion we draw from this brief study of the literature is that a rule-based system does provide many benefits in the short term. But for a real surveillance system, where the complexity of the rules which govern the scene is higher than we have so far considered, a more flexible approach could be taken. Ideally, one would like to use a fully probabilistic method, such as a Bayes Net.

6.3 The general reasoning process

We noted above that the analysis of visually simple, static scenes by Brand and Cooper was achieved by augmenting a rule-base with expert information until satisfactory explanations of the static equilibrium is achieved. This system depended, principally, on:

(a) Detection of the distinct objects as input into a reasoning process,

(b) A set of causal rules (e.g. knowledge of what a counterbalance is and why that may prevent the blocks toppling, as shown in Figure 6.1).

Inspired by their work, we observe that, using the low-level vision techniques we have developed, it is now possible to obtain the qualitative descriptions necessary. This information extraction is

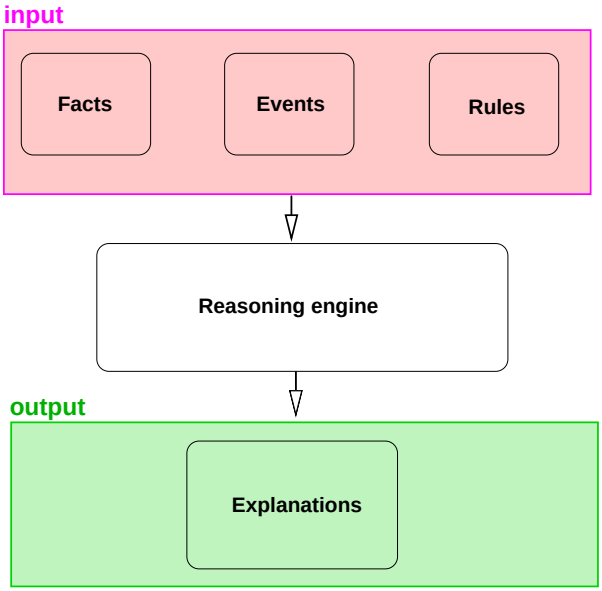


Figure 6.2: This schematic describes the basic reasoning process we employ in this chapter. For each different scenario, the reasoning engine is initiated when certain pre-specified events occur. The set of facts is searched for support for explanations generated using the rules, which are also pre-specified.

analagous to the detection of distinct objects, in the static case. With suitable expert knowledge of a dynamic scene, it is possible to extend the approach of Brand and Cooper to surveillance video and, indeed, human activity explanation in general.

Therefore, the general process we use for causal reasoning on the basis of such information is to specify a set of events which, when they are observed, trigger a search through the predefined rules and current facts list for the best explanation of the current activity. The events and the rules are interchangeable between scenarios. The reasoning engine is specified directly below in Algorithm 2 and in Figure 6.2.

Note that, while in this case, we do specify the events which require explanation, this is not always necessary. It would be possible to require that "unusual" events initiate the reasoning engine, where unusual is defined in relation to a threshold on the likelihood of the observed activity. Given that unusual activity could take any form, reasoning about it would require a more sophisticated (and much larger) system than that which we develop here, however. Hence we specify the events to be explained.

Algorithm 2 Reasoning process

```
load events-list
load rules
check facts for event in events-list
for all frames in sequence do
    update facts list
    if event occurs then
        derive hypotheses from the rule-set
        for all hypotheses do
            search known facts for hypothesis support
        end for
    end if
end for
```

6.4 Analysis of tennis play

6.4.1 Types of tennis players

Tennis is the first application on which we demonstrate causal reasoning. Tennis is a game with laws which bound the playing of the game. These laws can be used to encode a set of rules for causal reasoning. For example, a point must start with the shot known as a *service* which has a specific, well-defined form. This is one type of rule which could be encoded in a reasoning system.

Other types of rules are more high-level and relate to the way in which a player executes a game-plan. This typically involves a compromise between defending against the opponent's strengths and playing to one's own strengths. For example, a player may be a dedicated serve-and-volleyer. If this is indeed his/her preferred style of play then it will be observed that they attempt to manipulate the play in such a way as to enable them to play at the net.

How would one observe a player executing a serve-and-volley game plan? One example may comprise a player forcing the opponent to play a shot which takes long enough to return to provide enough time him/her to run to the net. A *lob* would be such a shot, and, more commonly, so would a regular forehand *deep* and *wide* to the baseline forcing the opponent to play a weaker shot at a limited angle. An experienced tennis viewer, player or a coach would have no trouble identifying a causal link in this chain of events. It could, in reality, be summed up by a commentator saying something like, "Henman forced Coria to retreat which allowed

him to charge to the net and play the winner". This is clearly an expert's summary of the point but contains all the evidence of a causal reasoning process. This is exactly the kind of process we demonstrate the ability to model.

6.4.2 Causal reasoning in tennis matches

Trigger event	**Corresponding rule**
transition to net	Move-to-Net
transition to baseline/backcourt	Move-to-baseline/backcourt

Figure 6.3: The predefined events and rules for reasoning about tennis matches. The rule corresponding to each event is initiated by the reasoning engine when an event is detected.

Causality is very clear in a sport such as tennis. That is, a player's actions are directly influenced by his opponent's actions and *vice versa*. Consider the most elementary causal relation in a tennis match: a player runs to the right-hand side of the court *because* his opponent played a shot to that region. This may seem elementary, but it is the essential element in constructing a convincing explanation of the events in a match.

The actions of a player can also be dictated by expert, higher-level knowledge. For example, if a player is injured, the opponent can exploit his resultant lack of mobility by playing shots which make the player run extensively around the court. To begin with, however, we start at the more basic level and pose a question that requires an answer based on causal relations:

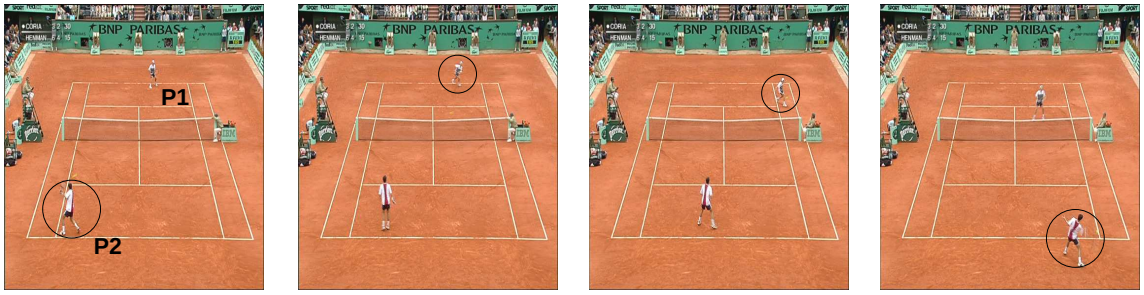
Why did the player run to the net?

There are two basic causal explanations which answer this question:

(a) The opponent forced him to the net,

(b) The player himself engineered the opportunity.

Formal rules which allow us to differentiate between the two explanations are therefore:



Frame	Explanation
66-95	Player 1 forced to run-to-midcourt
96-135	Player 1 initiated run-to-net

Figure 6.4: In this play, the player at the top of the image runs to the midcourt due to a shot played short by his opponent. He then decides to run to the net of his own accord. The automatic explanation is generated by a reasoning engine based on rules. The player with the ball in his court is circled at each frame.

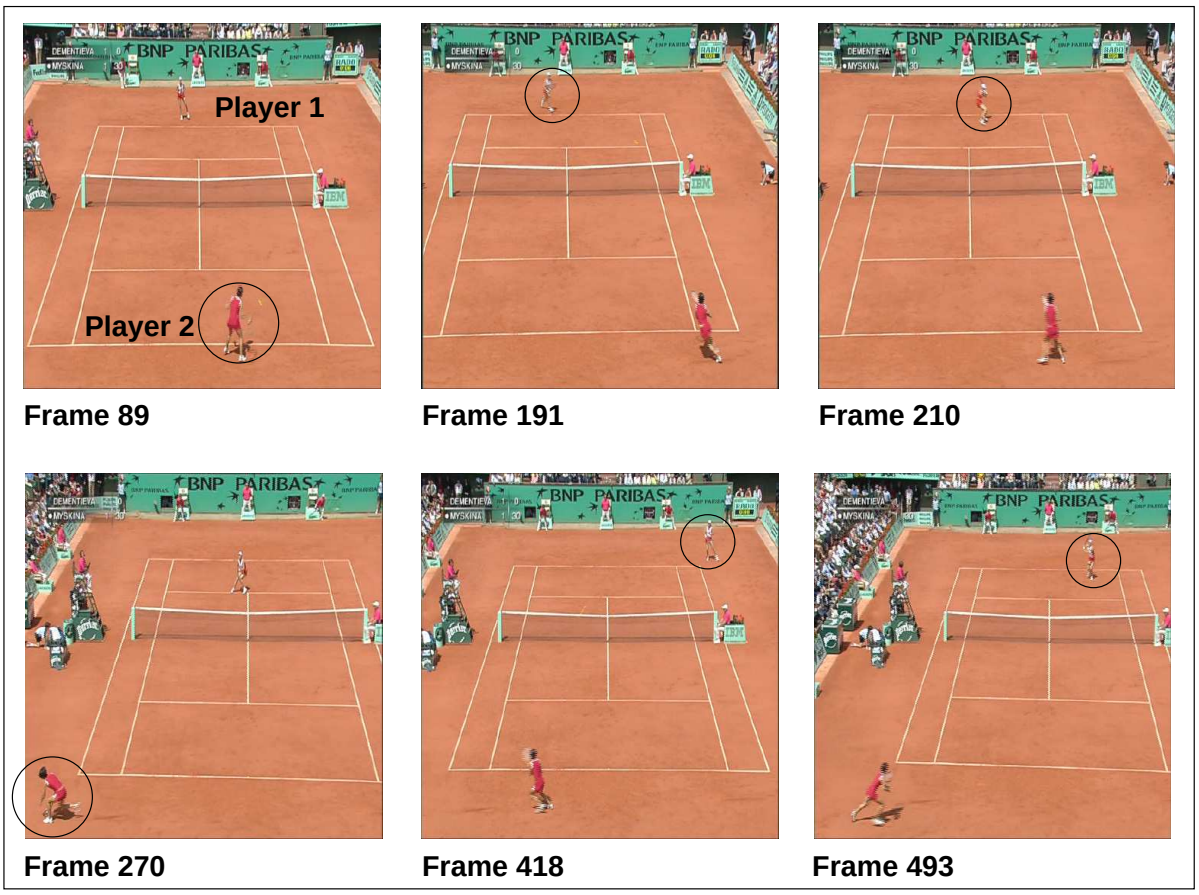
(a) IF the ball is played short by the opponent, THEN the player ran to the net *because* the opponent forced him to,

(b) IF the ball is played by the player and, during the non-shot period, he runs to the net, THEN the player *engineered* the opportunity to run to the net.

6.4.3 Generating causal explanations using rules

The input to the reasoning process is a set of qualitative facts, drawn from the action/behaviour/gaze estimates. Here, we take the ML likelihood result at every time step, although we consider in section 6.6.1 how to extend the reasoning process to be fully Bayesian. For every scenario, be that tennis or urban surveillance, the information available to the engine and the rules the engine uses will be different. In this tennis case-study, the facts contain qualitative text descriptions which are obtained automatically from the action and behaviour recognition stages of our system. These are as follows:

- Player 1 qualitative position,
- Player 2 qualitative position,



Frame	Explanation
89	Player 2 forced to retreat to backcourt
191	Player 1 forced to advance to baseline
210	Player 1 forced to advance to midcourt
270	Player 2 forced to retreat to backcourt
418	Player 1 forced to retreat to backcourt
493	Player 1 forced to advance to midcourt

Figure 6.5: Causal explanation of significant events during a point in a tennis match. The player with the ball is circled at each frame.

- Which player is currently playing a shot (derived from the proximity of the ball to each player).

These facts are analysed per frame and the reasoning process is initiated when certain data are observed corresponding to certain events e.g. a transition to the net. So, the corresponding rule "move-to-net", is initiated when a transition from qualitative positions, baseline to midcourt to net is observed. This rule is described at a high-level in software in terms of the qualitative information which is generated by the lower-level video analysis tools. An example is shown in Figure 6.4. This particular rule formally encodes the expert knowledge that a player can either respond to the opposition's shot or take initiative and move during the period when the ball is in the opponent's court.

Using the input facts, the rule can be encoded very efficiently as:

Algorithm 3 Tennis move-to-net rule

```
1: input facts
2: scenario = move to net;
3: player = facts.player
4: transitionTime = facts.transitionTime
5: if facts.currentPlayer(transition time) ≠ player then
6:    explanation = ["Player initiates" scenario "at time" transitionTime]
7: else
8:    explanation is ["Player forced to" scenario "at time" transitionTime]
9: end if
10: return explanation
```

An example of the reasoning process operating in response to the detected event is shown in Figure 6.4. In this example, it can be observed that the automatically generated description of this event contains causal information: the explanation is that Player 1 *initiated* (i.e. caused) the advance to the net, not that he was forced to by Player 2.

By extending the knowledge of the system to recognise general transitions between qualitative positions, the set of possible explanations is augmented to enable the system to generate explanations about transitions on the court in general. This is done by augmenting the event and rule-set to include expert knowledge about other transitions e.g. baseline to backcourt etc. A complete description of all such events for a second example is shown in Figure 6.5. This

demonstrates a high-level causal explanation of that particular tennis point and is a considerable extension of the commentaries we generated in previous chapters.

6.5 Rule-based agent behaviour analysis in an urban surveillance context

It will be observed that the tennis example we have discussed above did not make full use of either spatio-temporal actions, behaviour or gaze-direction information. In fact, the causality in tennis is so well-bounded that much can be achieved using positional information alone, especially when expert knowledge, with reference to player-types, is allied to the rules of the game. The clear nature of the causal relations enabled a straightforward encoding of rules to be used in a reasoning engine. What if we include all the information at our disposal using the tools this thesis has developed? How general is the rule-based approach in that case? To answer this question, we now apply the techniques to human interactions in an urban setting. This scenario has the advantage that gaze-direction is significant. Spatio-temporal action and overall individual behaviour must also be included to make sense of interactions within the scene.

As in the tennis examples, where the tennis player had partial knowledge of his opponent, we make the following assumptions about the person in an urban setting in order to formulate a person "agent":

- The agent has knowledge of his own state which includes action, behaviour, and gaze-direction,
- The agent can see other agents at a distance if they fall within the visual field (see Figure 6.6),
- The agent can sense anything within a specified range which is shorter than the visual field and reflects ability to, for example, hear someone walking behind,
- Interactions are possible within certain proximity, e.g. meeting.

Figure 6.6: The visual field of an agent is represented by the arcs highlighted in the image. See chapter 3 for details on how this region is estimated automatically.

In this model, an agent can only sense local information about another agent: spatio-temporal action and gaze direction. An agent does not know about other agents' goals or longer-term behaviour. Although the vision process which provides the input to the reasoning engine has knowledge of all activity in the scene as a whole, the agent, and thus the reasoning engine, is deliberately provided limited information. There is thus a set of facts associated with each agent, representing its entire knowledge. These facts are updated continuously and made available to the reasoning engine, but there is no "all-seeing" reasoning process taking place.

At each time step the following set of facts is updated:

1. The action, behaviour and gaze-direction of each agent;
2. The relative proximity agents (measured by the absolute Euclidean distance between the estimated centre of each agent);
3. The visibility of each agent to one another i.e. can Person 1 *see* Person 2?;
4. The relative directional headings between agents;
5. The directional headings of agents, individually.

Trigger events	Rules list
Move to road	Potential meeting
Move to pavement	Meeting
Move to drive	Ignoring
Stopped	Avoiding
	Together
	Proximity

Figure 6.7: The predefined events and rules for reasoning about interactions in an urban context. In this case, the reasoning is more complex than tennis, and so the rules are not directly initiated but various conditions must be met to fire rules. This is discussed in more detail in the text.

6.5.1 Detecting and classifying interactions between two agents using rules

People meeting with one another is a common occurrence in an urban scene. In fact, recognising groups of people versus independent individuals and, in particular, detecting cooperating individuals, is a core element of the human interpretation of urban scenes. Police surveillance officers, for example, may be interested in an exchange of illegal substances at a meeting of two individuals under observation.

There are many cues humans use to distinguish between people meeting or people ignoring one another. One such cue, discussed in chapter 3, is that people who are together will generally acknowledge each others presence by *looking* at one another periodically and at regular intervals. Other, more obvious cues include proximity. Here, we demonstrate using the same reasoning engine as was used for the tennis example to detect and classify such interactions. Therefore, by defining precisely what is required for the event "meeting" to take place we can distinguish between people passing one another and people meeting together. The set of trigger events and rules which can be initiated is shown in Figure 6.7.

Initially, the proximity of the individuals is analysed as shown in Algorithm 4, below.

First, a "potential-meeting" is identified when agents are within a predefined proximity for a predefined period of time (typically 100 frames) and also within one another's field of view.

The rule for meeting is that the intermediate state potential-meeting must be the current explanation of the interaction. Additionally, the agents must be performing the same spatio-temporal action e.g. they are both *walking-on-the-pavement*:

Algorithm 4 proximity rule

```
load facts
proximityThreshold = 100
timeThreshold = 100
for all frames do
    distance = (P1 position) - (P2 position)
    if distance ≤ proximityThreshold then
        if p1action = p2action & P1 visible & P2 visible then
            together = 1
            increment = increment + 1
        end if
    end if
    if increment ≥ timeThresh then
        situation = "together"
    else
        situation = "not together"
    end if
    update facts
end for
```

By contrast, an "ignore" rule is initiated when the conditions for meeting are not met but when a "potential-meeting" has previously occurred. If none of these agent states are identified, there is no interaction defined.

These rules can be encoded as shown in Algorithm 5.

Figures 6.8 and 6.9 show these rules in operation in the urban surveillance context. Algorithm 5 explicitly defines the rule for the scenario "meeting".

6.5.2 Rule-based causal reasoning

In the examples of Figure 6.8 and 6.9 there are events consisting of two independent agents interacting. The *reasons* for these events occurring are not apparent directly from the video. That is, the person in Figure 6.8 who crossed the road in order to meet his friend may have done so *because* it was pre-arranged or *because* he happened, by coincidence, to see him. It is not possible to distinguish between these postulated, hypothetical reasons from the data alone. This is still true even if the scene rules are completely known, as it requires detailed knowledge of the intention, goals and history of a specific individual, which is not reliably available, certainly not in a general surveillance application where the individuals under observation are generally

Figure 6.8: (*Clockwise from top-left*) The `Meeting` rule is initiated in this case.

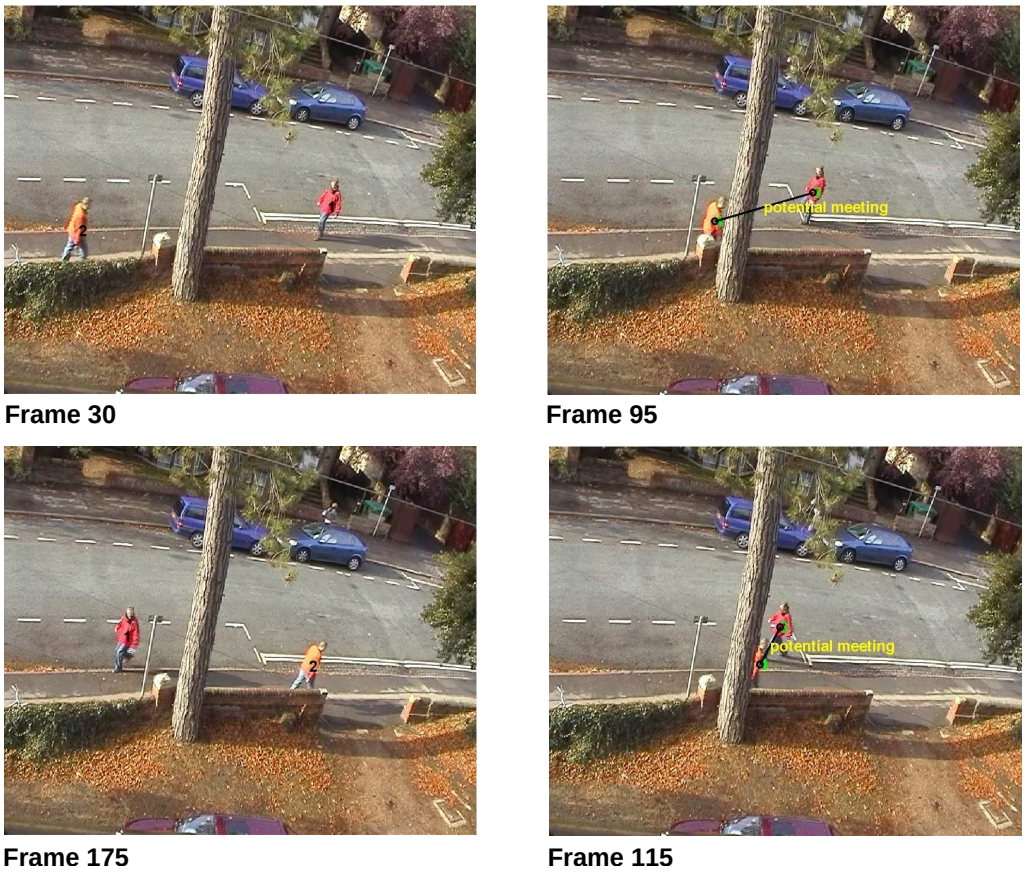


Figure 6.9: (*Clockwise from top-left*) The `Ignore` rule is initiated after the `Potential-meeting` rule.

Algorithm 5 meeting rule

```
load facts
meetingThresh = 50
j=lastFrameIndex
for i = 1 to j do
   if situation(i) = situation(i-1) then
      if situation(i) = "together" then
         togetherInc = togetherInc + 1
      else
         togetherInc = 0
      end if
   end if
   if togetherInc ≥ meetingThresh then
      scenario = "meeting"
   else if togetherInc < meetingThresh & togetherInc > 0 then
      scenario = "potential meeting"
   else
      scenario = "not meeting"
   end if
   update facts
end for
```

unknown.

A lower-level of causality is still in operation. In fact, that can be inferred in the sentence above. It was stated that the person, "...crossed the road *in order to* meet ...". This type of causality is amenable to analysis using the information we currently can obtain about both the scene and the agents. For example, the question could reasonably be posed: "Why did the person walk onto the road?". The causal explanation, at this lower-level, would be that he did so *in order to* meet his acquaintance.

In the particular urban scene of Figures 6.8 and 6.9, there are a number of events which can occur which could be explained in terms of causal relations. The examples we have shown suggest that transitions in qualitative action are of interest and can, subsequently, generate interesting activity. The overall reasoning process is therefore identical to that used in the tennis example, with a different set of events and rules.

At each frame of the input footage, as the activity is estimated directly from the video, transitions between actions are searched for, triggering an "event" which requires to be explained.

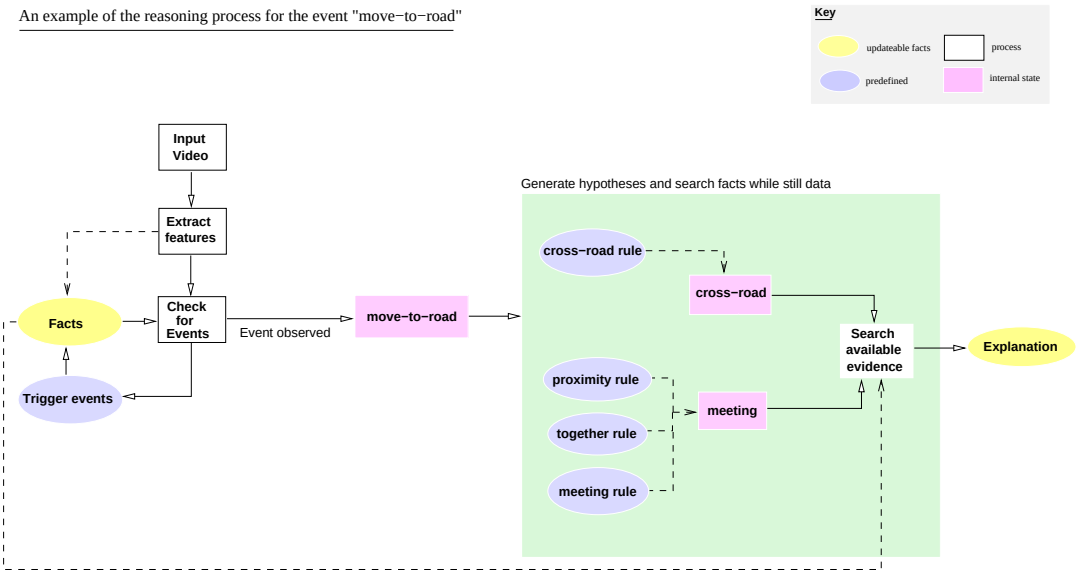


Figure 6.10: A schematic diagram of the reasoning process initiated when the event "move-to-road" is detected.

The facts are then analysed searching for support for particular hypotheses which could explain the event sequence. Therefore, in the example of Figure 6.8, the transition between qualitative actions *walking-on-farside-pavement* and *walking-on-road* generates an event "move-to-road". This event essentially poses the question, "Why is the agent now walking on the road?". A graphical illustration of the overall reasoning process for answering this question is shown in Figure 6.10.

Hypotheses to explain this particular scenario are defined, using human knowledge of this environment, as:

1. IF the event "move-to-road" is followed by event "move-to-pavement" AND the current location is not the same as the location triggering the first event (i.e. the road is crossed) AND, subsequently, a meeting takes place THEN the explanation is that, "the agent crossed the road to meet the other agent",

2. IF a crossing of the road is observed NOT followed by an interaction THEN the explanation is that the agent crossed the road,

3. IF a "move-to-road" event is triggered AND subsequently a "move-to-pavement" event but back to the same pavement THEN no explanation is provided UNLESS another agent

was in the near vicinity THEN the explanation is that it was necessary to avoid collision.

The pseudo-code for this scenario is shown in Algorithm 6.

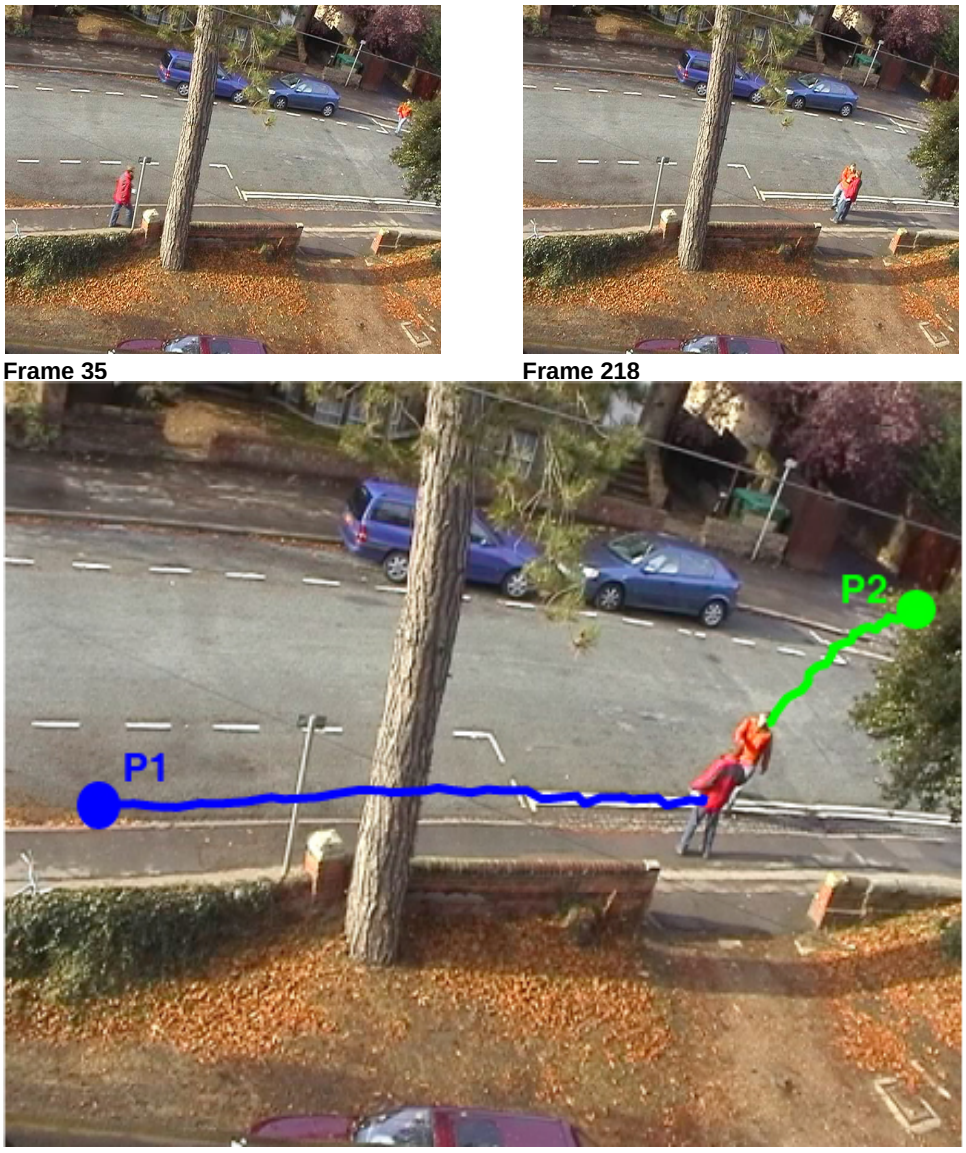
Algorithm 6 move-to-road rule

```
load facts
if event="meeting" then
   for j = 1 to lastFrame do
      if scenario = "meeting" then
         currentAction = facts.positionLabel(j)
         explanation = "Person" event "to meet on" currentAction
      end if
   end for
   for j = 1 to lastFrame do
      if scenario = "ignore" then
         currentAction = facts.positionLabel(j)
         explanation = "Person" event "to avoid other Person on" currentAction
      end if
   end for
end if
```

Similarly, we generate hypotheses for explaining events such as "stopping", "move-to-pavement" and "move-to-driveway".

It can be seen, from the rules itemised above for the particular event "move-to-road", that these rules are: (a) general to all such urban scenes, and (b) easily augmented. For example, in the case of the 3rd rule, it is conceivable that this behaviour could be observed when a car is passing and an agent accidentally steps out before looking to see the car. Clearly, if we were tracking cars in addition to people, and trying to explain their role in the scene too, extending the rule set is simple to deal with this new hypothesis.

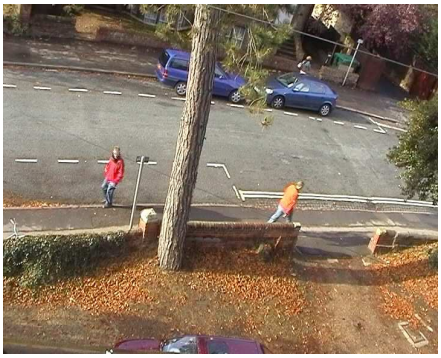
The best explanations of the observed activity are generated using the reasoning engine. In Figures 6.11 and 6.12 the output for two different situations which is automatically generated by our system, is shown. These results represent true causal reasoning about interesting human activity in video.



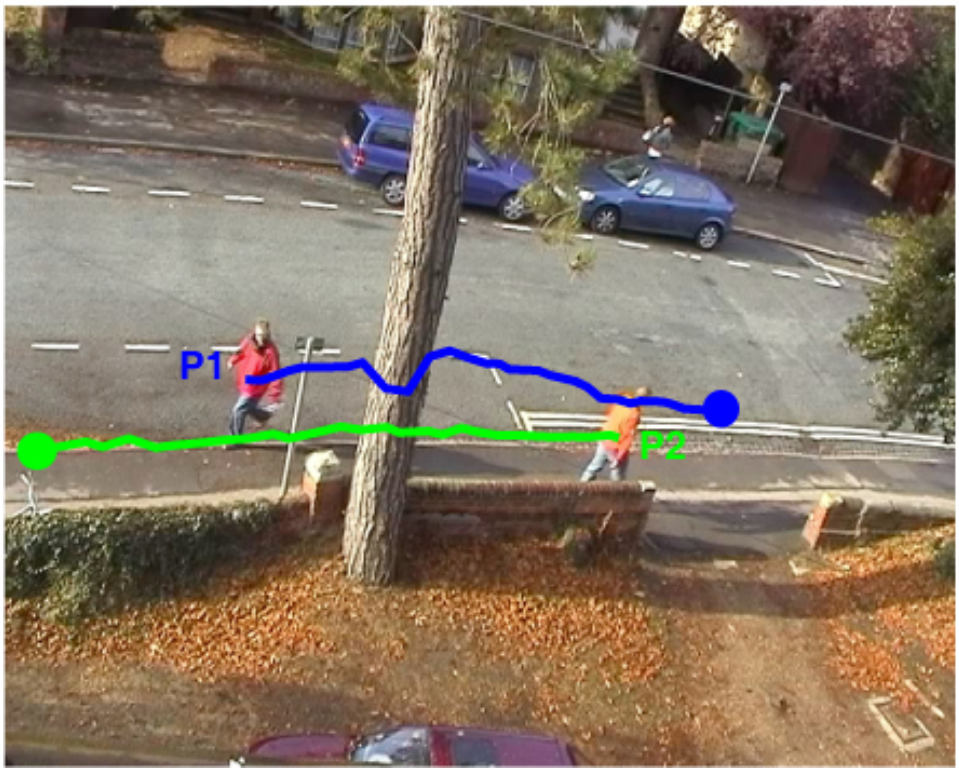
Frame	Event explanation
35	Person 2 move-to-road to Meet on nearside-pavement
218	Person 2 move-to-pavement to Get-off-road

Figure 6.11: In this example, Person 2 is under observation. Two significant events are observed. The first (*left*) is leaving the pavement to walk on the road. The second (*right*), leaving the road to walk on the opposite pavement. By searching the action/behaviour/gaze data for an explanation for these events the causal explanations below the figures are automatically generated from known rules of this scene. Note that the estimated gaze-directions for this sequence can be seen in Figure 3.16.

Frame 60



Frame 138



Frame	Event explanation
60	Person 1 move-to-road to Avoid Person 2 on nearside-pavement
138	Person 1 move-to-pavement to Get-off-road

Figure 6.12: In this example, the event where Person 1 steps on to the pavement requires explanation. The best explanation is that the person avoids the other.

6.6 Conclusion

6.6.1 Future work

We noted above that it would be desirable to use all of the probabilistic information available to reason about the activity. In the work described in this chapter, we have used ML descriptions arising from the action-recognition stages of our system. As a step towards a fully Bayesian causal reasoning method, we can begin by modelling two types of tennis player using extended HMM state transitions, compared to those of section 5.12. An HMM could easily take distributions over action (e.g. multi-variate Gaussian) as input/output. Expert knowledge about two player-types is summarised below.

The Serve-and-volley player:

- Desire: move to net after his own first shot,
- Intention: limit opponent's options and force weaker shot.

The Baseliner player:

- Desire: stay at baseline,
- Intention: open up court by forcing opponent wide or short.

Using these basic definitions, we can define a state transition matrix for each player. The states include the player's own position and the position of the opponent. These are shown in Figure 6.13. Note that for the type of reasoning we are modelling, the position of the ball is required (which is hand-tracked for the experiments which follow). For instance, the player runs to where the ball has been played, regardless of the shot type, reacting to the opponent's shots. However, a more sophisticated reasoning engine may simulate the *predictions* of a player based on the agent's belief that the opponent is about to play a certain shot.

The transition matrices shown in Figure 6.13 have been written down by hand. They represent the first level of causal reasoning. Take, for example, the baseliner player model (Figure

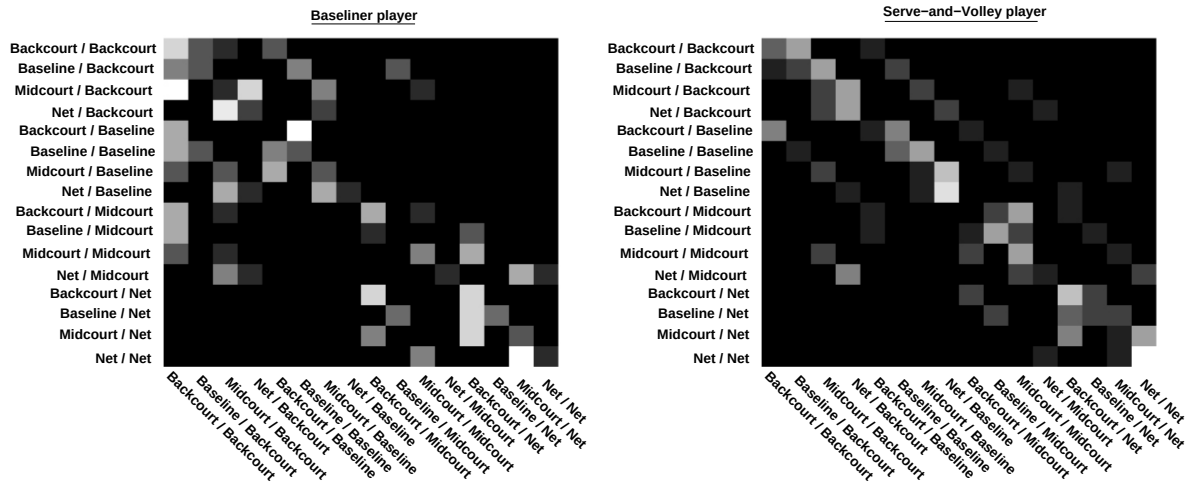


Figure 6.13: The state transitions for the smoothing HMMs for two types of tennis agent, Baseliner (*left*) and Serve-and-Volleyer (*right*). Each state in the transition matric represents the known position of the agent and the observed position of the opponent e.g. *baseline / net* is the state when the agent is at the baseline but the opponent is at the net. Thus, decisions which are dependent on the opponent can be encoded. For example, the Serve-and-Volleyer transition matrix (*right*) encodes the belief that if the opponent is observed to be at the net, the player (despite having a preference for playing at the net himself) will be most likely to stay at the backcourt/baseline.

6.13, *left*). The transition matrix encodes the preference to remain at the baseline/backcourt regardless of the opponent's position. But if the baseliner finds himself at the net, the transition matrix tells us that he prefers to stay there. By creating states which represent the position of both the player under observation and the opponent, the transition matrix encodes the probability of positional changes conditioned on where the opponent is observed to be at that time.

Results of modelling players according to these types are shown in Figures 6.14 and 6.15. One can see that the likelihood of the HMM explaining the observations could provide an indication as to: (a) whether the model is correct i.e. the player really is of the type supposed or, more interestingly, (b) which player is dominating the point, assuming the models are correct.

6.6.2 Comments

The work of this chapter has exploited the results of the previous chapters to demonstrate that efficient, rule-based, causal reasoning can be implemented when, (a) expert knowledge of the rules which govern human activity within the scene is available, and (b) the information required

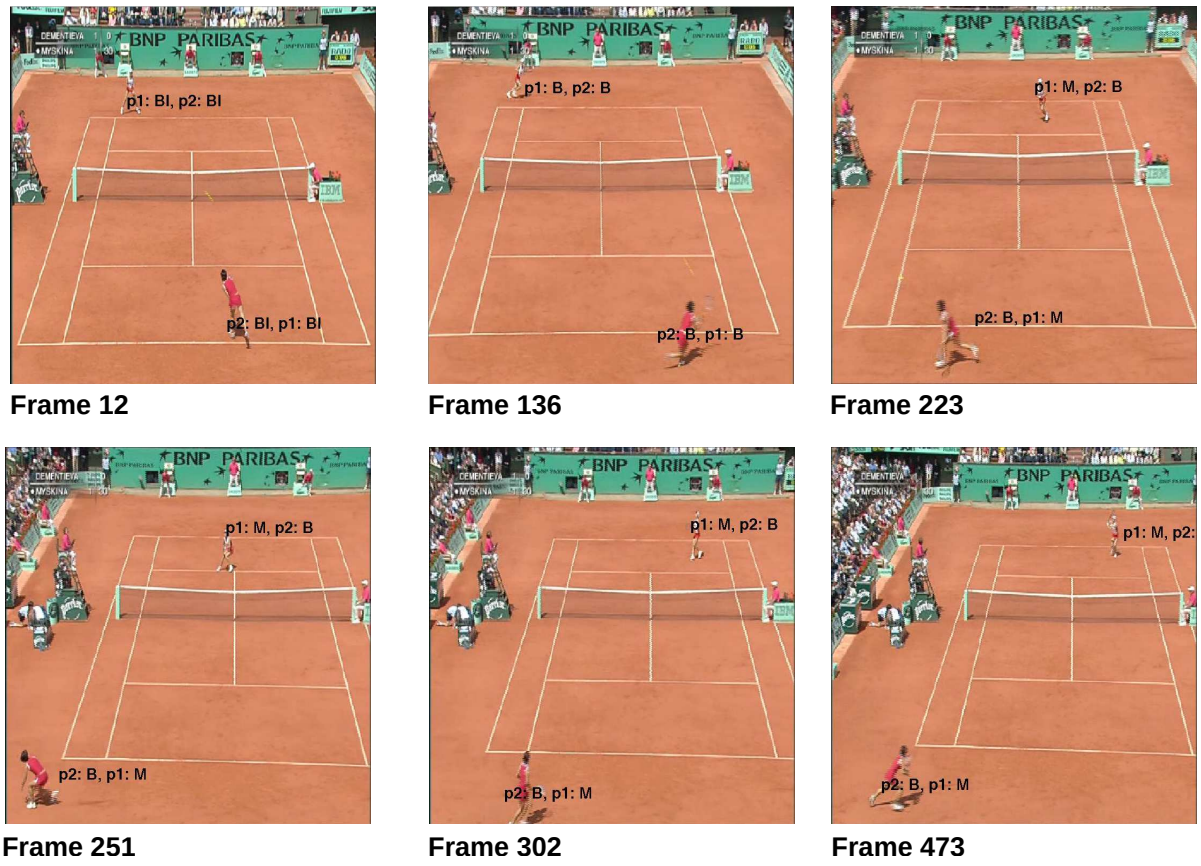


Figure 6.14: The most likely (joint position) state for each player is computed for the raw shot sequence which is smoothed using the *Baseliner* HMM of Figure 6.13, which is correct, judging by the play seen here. For selected frames, the ML (Viterbi) state is shown in this figure. The state labels are superimposed beside each player and the key is: B=Backcourt, Bl=baseline, M=midcourt, N=net.

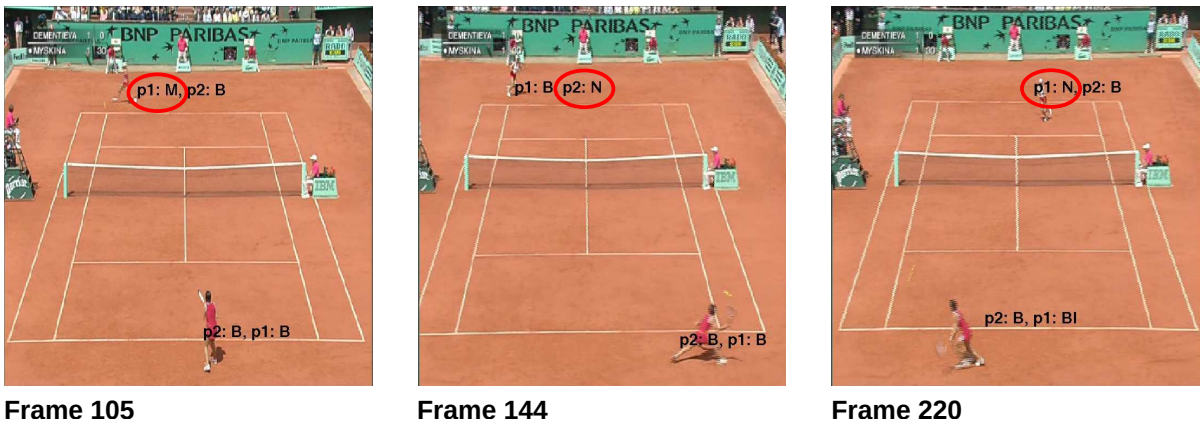


Figure 6.15: In contrast to the results of Figure 6.14, when the smoothing model is incorrectly chosen, the state estimate errors can be clearly seen. In this example player 1 (top in each image) is modelled as a *Serve-volleyer*, for which there is little evidence in this play.

to reason at a higher-level can be extracted directly from video i.e. qualitative descriptions of activity.

We demonstrated the use of rules to automatically generate causal descriptions of activity in two scenarios: tennis and urban surveillance. By requiring, in both scenarios, that it is certain predefined events which require explanation, we automatically searched the known rule set and generated causal explanations of human activity in video, when those events occurred.

The generality of the technique is highlighted by the fact that the same reasoning engine is used for completely different scenarios with only the set of rules and trigger events being interchanged for each application.

The main drawback is that hard decisions about activity are made. Moreover, the inputs to the reasoning process are the ML estimates of each action/behaviour/gaze estimate, not the actual probability distributions. This latter point strongly suggests that further work can be done to fully exploit the probabilistic estimates, such as using a full Bayes Net for inference on causal relations. Indeed the work of Pearl on Causality could be particularly useful in this context [111].

One particular limitation of not having a fully probabilistic representation is that *prediction* is not truly possible. This is because, without the likelihoods associated to current observations, distinguishing between the quality of evidence for competing hypotheses is not possible in any meaningful way. This is a clear line of future work, as we have begun to show for a tennis scenario in section 6.6.1 of this chapter.

7

Conclusion

This thesis ends with a brief recapitulation of the topics we have discussed and a list of the novel contributions of this work. We conclude with a discussion on future research directions to utilise and extend the results of this work.

7.1 Summary of the thesis

This goal of this thesis was to, *Develop a set of techniques to enable automatic causal reasoning about human activity as recorded in surveillance video.*

We began by posing the question, "What information does an expert require to reason about human activity?" Looking at the prior work in the published literature it was apparent at an early stage that many of the current techniques which would naturally be used to generate the required low-level estimates of human activity are not appropriate for the resolution of the data in surveillance video.

So, in chapter 3 we developed a novel technique for estimating head-pose in images where the head images are low-resolution. This estimate was refined by contextual information in a Bayesian fashion to produce distributions over potential gaze directions.

Gaze direction is an important cue to the intention of a person, but not the only piece of significant information required to reason about, or report on, their activity. Using a similar approach, defining a descriptor which is readily computed from low-resolution video and interpreting in a probabilistic fashion, we developed a technique for general analysis of human activity in video. Our approach was motivated by the fact that, in genuine surveillance operations, the analysts have strong domain knowledge. As such, we used training data which had been hand-annotated, and generated probabilistic samples over the training data for new examples in chapter 4. The independent features were combined, using Bayesian probability theory, to generate estimates over all spatio-temporal activities in the training data for the new example.

In chapter 5, recognising that the behaviour of a person over time is composed of a chain of spatio-temporal actions, we encoded scene "rules" as a set of state transitions in a Markov Chain where the states are indices into actions with their respective likelihoods. This enabled faster generation of higher-level behaviour analysis tools without discarding the benefits of compact models. These behaviour HMMs are also general to the scene itself, not a specific viewpoint of the scene.

Having achieved the extraction of low-level, probabilistic information about human activity, we demonstrated in chapter 6 that this information can be used to reason, causally, about interactions between people and generate realistic, high-level explanations of overall activity. This is, to the best of our knowledge, the first demonstration of causal reasoning about human activity in surveillance video which operates directly from the video stream, with minimal manual intervention.

Thus, we verified our *thesis*, which is: *In order for a computer system to effectively, and automatically, reason about human activity in surveillance video, low-level vision techniques must first abstract the information a human would require, from the video, to an intermediate, probabilistic and qualitative representation based on motion.*

7.2 Contributions

Human activity recognition is very much at the fore-front of current Computer Vision research. Surprisingly few papers in the published literature address the problem of generating descriptions of human motion in video using only the behaviour and not the appearance of the person. This is despite the overwhelming evidence from vision psychology that motion is the primary cue necessary for interpreting causality in visual data. Moreover, considerations of true operational conditions and how to effectively utilise the expertise of the "man-in-the-loop" have been all too rare in the development of low-level techniques for human activity recognition. This thesis has contributed to each of these important areas. We list the contributions as they appear:

- **Robust estimation of gaze-direction from low-resolution faces.** We developed a new technique for estimating gaze-direction in surveillance-style video footage, in contrast to the well-studied problem of estimating gaze in high-resolution (HCI) video. Our method combined a head-pose descriptor based on skin detection with body-direction and was demonstrated primarily on footage from the surveillance domain including standard vision sequences. This work was published in [130, 132]

- **Application of action-recognition at a distance techniques for use in a surveillance system and automatic sports commentator.** We extended an existing technique for recognising person-centred activity at the medium/low resolution, to generate probabilistic distributions of action using the novel application of a fast database sampling method. Additional distributions of position and velocity features, fused with the person-centred action distribution, using a Bayes Net, generated distributions over spatio-temporal actions. The ML spatio-temporal action estimate at each time step represents a robust commentary on sports or surveillance video. This work was published in [131], and is currently in press [133].

- **The development of a general framework for exploiting expert prior knowledge.** Recognising that behaviour can be composed of sequences of action, we developed a new framework for behaviour recognition. This technique used stochastic models which encode the rules for action sequences corresponding to behaviours extended in time. These models have as their input/output the ML spatio-temporal action index and associated likelihood and are therefore general to the scene. This framework allows expert knowledge, where available, to be rapidly incorporated while retaining the benefit of compact stochastic models. This work was published in [131, 134], and is currently in press [133].

- **Causal reasoning about human activity directly from video.** The robust extraction of action, behaviour and gaze information enabled us to demonstrate causal reasoning about human activity directly from surveillance and sports video. This is the first demonstration such a reasoning process. Previous attempts from the AI community suffered from a lack of robust Computer Vision techniques, whereas vision literature has generally focussed on low-level information. This work was published in [134].

This thesis has demonstrated automatic causal reasoning about human activity in video, and this research field is wide open, both in terms of the basic research required to make intelligent surveillance a reality, and in the huge variety of applications which require robust visual surveillance methods. The work we have presented has pointed, we believe, in the correct direction for vision researchers who are seeking to develop visual surveillance systems. There are a number

of interesting problems which we would suggest require consideration when augmenting or using our results. We now discuss some of these, briefly, in conclusion.

7.3 Future research directions

7.3.1 Defining surveillance ontologies

The set of descriptions which were specified in each of the application domains on which we demonstrated our methods, were defined by a person with detailed, but not professional, knowledge of the scene. In the future, due to the nature of the funding for this work, we expect that the results of this thesis will be be exploited to achieve semi-automatic reporting of surveillance footage for Royal Air Force imagery analysts. These analysts are trained to use a well-structured and clearly-defined language when describing what has been observed and filing reports on the activity they have been tasked to survey.

A researcher's "ontology" (such as has been employed in this work) for describing human activity will not, generally, seem realistic to a true expert in the domain. In the military or law-enforcement context the researcher's descriptive language may well be misunderstood. This could create significant problems within the chain-of-command leading to operational failure. Therefore, it is critical that researchers seeking to develop systems involve the user in this phase of development to avoid ambiguity at all costs.

Moreover, an interesting research topic will be to explore the possibility of defining a robust surveillance ontology for general use in urban environments.

7.3.2 Extracting further low-level information from video

The development of a genuine ontology will directly influence the information which is required to be extracted from video. For example, it may be that the height of an individual is required to be known. In which case, multiple-camera methods could be implemented to extract 3-D information from the scene, which, when allied to some knowledge of metrics within the scene (e.g. the height of a lamppost) this information can be obtained. This is an example where

known techniques can be exploited to augment the agent's knowledge about a scene.

However, it is also very likely that new computer vision methods will need to be developed to obtain information. For example, an area we have not considered in this thesis is night-time surveillance. How far the methods demonstrated in this thesis apply to night-vision or infra-red imagery is not yet clear. This is one area which, when studied, could yield new and interesting vision algorithms and techniques.

7.3.3 A fully probabilistic reasoning system

The most obvious future direction arising from this work is the incorporation, and extension, of these results into a fully probabilistic reasoning framework. It will reasonably be asked why we spent considerable effort maintaining probabilities throughout only to discard this information when it was, arguably, most useful. For reasons of expediency we used the ML estimates from probability distributions on a number of occasions, most significantly when developing a causal reasoning system.

We strongly suggest that the research and implementation of a full Bayesian Network, which can be interpreted as defining the causal relations between variables, would be the most natural extension and application of the novel techniques which have been generated in this thesis. This would be very much in keeping with the approach we have taken throughout and would yield the considerable benefit of providing a fully probabilistic interpretation of human activity to a user, leaving the final decision-making to the analyst.

7.3.4 The role of learning in reasoning systems

We noted in chapter 6 that causal reasoning failure can be an opportunity to learn. This is an area which would benefit from investigation. In particular, we have specified a set of rules which, in certain constrained scenarios will enable the system to arrive at a sensible conclusion. What happens when no satisfactory conclusion is reached? Methods for providing the system with the required knowledge at this failure point could be investigated.

A

Colour-based tracking in video

A.1 Mean-shift tracking in video

Figure A.1: Automatic initiation of targets (*left-to-right*) 1. Background image, 2. Foreground image, 3. Segmented object blobs, 4. Initiated targets for mean-shift tracking.

By tracking an object we achieve repeated measurement of the location of a moving target throughout the frames of a video. Tracking can be challenging due to the fact that the target may change in shape or appearance as the target orientation varies in relation to the camera. Additionally, there may be some small per-frame camera motion e.g. camera-shake due to wind or smooth panning by the operator to centre a moving target. Using colour alone to define the target provides invariance to shape changes so long as the appearance of the true target remains sufficiently different from background clutter.

The target of interest can be initiated by hand or by using background subtraction, as shown in Figure A.1, and the target model (histogram) thus defined. The mean-shift algorithm uses the Battacharyya coefficient as the similarity measure between two distributions which are discretised into u bins: $p(y)$ at the current image window centred at y and q, the target model histogram. This is given by:

$$\rho(p,q) = \sum_{u} \sqrt{p_u q_u} \tag{A.1}$$

which is maximised using an efficient iterative algorithm introduced by Comaniciu in [31]. Each pixel, x, in a window (centred on the current target location y_0) is assigned a weight:

$$w_x = \sum_{u} \delta[I(x) - n] \sqrt{q_u / p_u(y_0)} \tag{A.2}$$

The new estimate of the target position is computed as:

$$y_1 = \frac{\sum_x x w_x k(x,y)}{\sum_x w_x k(x,y)} \tag{A.3}$$

where k is a kernel which weights pixels close to the centre of the current window higher than those at the edge, in our case we use a Gaussian kernel (the Epanechnikov kernel - which is an approximation to the Gaussian - is also suggested). The iteration stops when $|y_1 - y_0| < \epsilon$, where ϵ is a predefined threshold, typically 1 pixel.

Search in scale-space is interleaved between each step of the gradient-descent in position (described above) i.e. a set of Gaussian kernels are defined with:

$$\{\sigma_s = \sigma_0 * b^s, -n \leq s \leq n\} \tag{A.4}$$

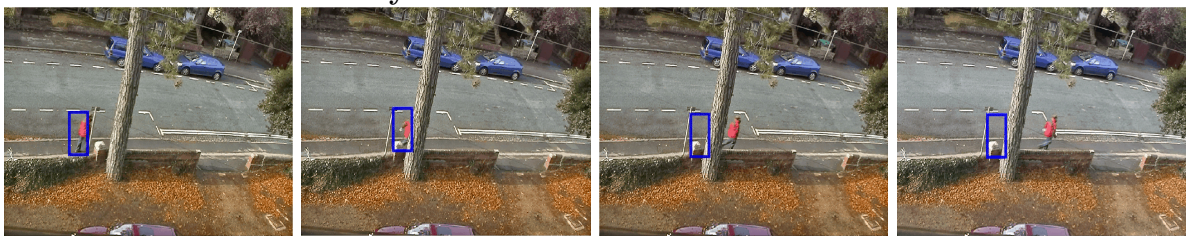
where $b > 1$ is the base of the logarithmic scale and n defines range of the search in scale around the current scale σ_0. We choose $b = 1.1$ and $n = 2$ as in [30]. The effect of tracking in scale, as well as image space, is shown in Figure A.2.

While the mean-shift algorithm as described here offers a degree of robustness to changes in target appearance (as shown in Figure A.3) it will, as with all simple vision-based tracking algorithms, fail where the target is completely occluded. In order to provide robustness to occlusion we implement the improvement of Bibby and Reid [10]. In their work, when the Battacharyya coefficient drops below a certain value, the search window is expanded by computing the Battacharyya coefficient for a grid of windows around the current location and, provided the target has not disappeared altogether or moved outwith even this wider search region, the location can be recovered. An example of the utility of this method in a surveillance context is shown below in Figure A.4.

Figure A.2: By tracking in scale-space as well as position the tracking is more robust as this example shows. Also it is preferable that as much of the background be eliminated from the target as possible when the target-centred image will be used for further processing.

Figure A.3: In this sequence two people meet and partially occlude one another. It is shown that the target tracking algorithm employed here (without additional occlusion reasoning) has a degree of robustness to partial occlusion.

Without occlusion recovery



With occlusion recovery

Figure A.4: (*First row*) The standard mean-shift algorithm fails when the target is completely occluded because the search window does not extend to the point where the target reappears. The position update has no better estimate than the current location since the true model has disappeared and, in general, all of the local background represents an equally dissimilar colour histogram compared to the target histogram. (*Second row*) By expanding the search window when the histogram similarity measure (the Battacharyya coefficient) falls below a specified threshold it is possible to recover the true target location.

B

Bayesian estimation: the Kalman Filter

B.1 General Bayesian estimation

In tracking we want to recursively estimate the state sequence of a target at a given time based on the set of available measurements. The state sequence at time k we write as $\mathbf{x}_k$, given by

$$\mathbf{x}_k = \mathbf{f}_k(\mathbf{x}_{k-1}, \mathbf{v}_{k-1}) \tag{B.1}$$

which is a function of the previous state and a noise process. The set of measurements is denoted $\mathbf{z}_{1:k} = \{\mathbf{z}_i, i = 1, \ldots, k\}$ where

$$\mathbf{z}_k = \mathbf{h}_k(\mathbf{x}_k, \mathbf{n}_k) \tag{B.2}$$

is a function of the previous state and some noise process.

B.1.1 Prediction step

If the distribution $p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})$ is available at time $k-1$ then the system model in B.1 is used to find the prior distribution for the next time step k by:

$$p(\mathbf{x}_k|\mathbf{z}_{1:k-1}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})d\mathbf{x}_{k-1} \tag{B.3}$$

Since the system model describes a first-order Markov process, the above equation has been simplified by the fact that $p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_{1:k-1}) = p(\mathbf{x}_k|\mathbf{x}_{k-1})$.

B.1.2 Update step

At time step k a measurement $\mathbf{z}_k$ is available, which is used to update the prior by Bayes' rule:

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) = \frac{p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{z}_{1:k-1})}{p(\mathbf{z}_k|\mathbf{z}_{1:k-1})} \tag{B.4}$$

which, to define these terms, says:

$$posterior = \frac{(prior)(likelihood)}{evidence} \tag{B.5}$$

B.1.3 Kalman Filter

With the Kalman Filter it is assumed that at each and every time step the posterior distribution is Gaussian. As a result it can be parameterised by a mean and covariance. It can be shown that if $p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})$ is Gaussian then $p(\mathbf{x}_k|\mathbf{z}_k)$ is also Gaussian if the following assumptions are true [68]:

- the system and measurement noise, $\mathbf{v}_{k-1}$ and $\mathbf{n}_k$, are drawn from Gaussian distributions with known parameters,
- the system and measurement models are known and are linear functions of state and noise.

That means we can write B.1 and B.2 as

$$\mathbf{x}_k = F_k\mathbf{x}_{k-1} + \mathbf{v}_{k-1} \tag{B.6}$$

$$\mathbf{z}_k = H_k\mathbf{x}_k + \mathbf{n}_k \tag{B.7}$$

These are linear functions, and the covariance of $\mathbf{v}_{k-1}$ and $\mathbf{n}_k$ are defined as the covariance matrices Q_{k-1} and R_k

The Kalman Filter algorithm (which is derived using B.3 and B.4) can be written as a recursive relationship:

$$p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1}) = \mathcal{N}(\mathbf{x}_{k-1}; m_{k-1|k-1}, P_{k-1|k-1}) \tag{B.8}$$

$$p(\mathbf{x}_k|\mathbf{z}_{1:k-1}) = \mathcal{N}(\mathbf{x}_k; m_{k|k-1}, P_{k|k-1}) \tag{B.9}$$

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) = \mathcal{N}(\mathbf{x}_k; m_{k|k}, P_{k|k}) \tag{B.10}$$

where $\mathcal{N}(x; m, P)$ is a Gaussian distribution with argument x, mean m and covariance P, and

$$m_{k|k-1} = F_k m_{k-1|k-1} \tag{B.11}$$

$$P_{k|k-1} = Q_{k-1} + F_k P_{k-1|k-1} F_k^T \tag{B.12}$$

$$m_{k|k} = m_{k|k-1} + K_k(\mathbf{z}_k - H_k m_{k|k-1}) \tag{B.13}$$

$$P_{k|k} = P_{k|k-1} - K_k H_k P_{k|k-1} \tag{B.14}$$

The covariance of the innovation term $\mathbf{z}_k - H_k m_{k|k-1}$ is

$$S_k = H_k P_{k|k-1} H_k^T + R_k \tag{B.15}$$

and the Kalman Gain is defined by

$$K_k = P_{k|k-1} H_T^k S_K^{-1} \tag{B.16}$$

The likelihood function of the model at time step k is calculated from a Gaussian distribution with a mean of zero, the covariance equal to the covariance of the innovation and the innovation as the argument. Hence the likelihood of the model is the sum of the likelihood function at each time step over all time steps.

This filter is optimal if the assumptions made at the start are true. This implies that no filter can do better than a Kalman filter in a linear, Gaussian situation. For problems which are clearly not linear and Gaussian approximations are necessary and form the basis of the Extended Kalman Filter, Unscented Kalman Filter and Particle Filters.

C

Expectation Maximisation

C.1 Learning state-space models

A Kalman filter is clearly a state-space model and hence the structure is known *a priori*. However, given a complete set of data, we would still require to learn the other parameters of the model as well as the hidden variables. If there is only one unknown it is possible to calculate the Maximum Likelihood directly by finding the probabilities of the unknown given the conditional variables and summing at each stage.

In this case the parameters and the hidden states are estimated by holding one fixed, maximising the likelihood with respect to the other and vice-versa. This is the basis of the Expectation-Maximisation (EM) algorithm.

Using any distribution Q over the hidden states, a lower bound on the likelihood L can be obtained [60]:

$$\begin{aligned}\log \sum_x P(z,x|\theta) &= \log \sum_x \frac{Q(x)P(z,x|\theta)}{Q(x)} \\ &\geq \sum_x Q(x) \log \frac{P(x,z|\theta)}{Q(x)} \\ &= \sum_x Q(x) \log P(x,z|\theta) - \sum_x Q(x) \log Q(x) \\ &= F(Q,\theta)\end{aligned} \tag{C.1}$$

(This equation is for a single observation z, hence the lack of subscripts.)

So EM alternates between maximising F with respect to Q and θ while holding the other fixed. It is initialised with some guess at the parameters.

C.1.1 E-step

The hidden variables are updated by

$$Q_{k+1} \leftarrow \arg\max_Q F(Q,\theta_k) \tag{C.2}$$

The maximum in the E-step results when $Q_{k+1}(x) = P(x|y,\theta_k)$.

C.1.2 M-step

$$\theta_{k+1} \leftarrow \arg\max_{\theta} \sum_{x} P(x|y,\theta_k) \log P(x,y|\theta) \tag{C.3}$$

The maximum of the M-step is calculated by maximising $\sum_x Q(x) \log P(x,z|\theta)$.

C.2 Using EM to learn state-space models

Since we have a first order Markov process, the log probability of the hidden states and observations for linear-Gaussian state-space models can be written as

$$\log P(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = \log P(\mathbf{x}_1) + \sum_{k=1}^{T} \log P(\mathbf{z}_k|\mathbf{x}_k) + \sum_{k=2}^{T} \log P(\mathbf{x}_k|\mathbf{x}_{k-1}) \tag{C.4}$$

Each of these probability densities is Gaussian, and so the overall expression is a sum of quadratics. So, given that $\mathbf{z}_k = \mathbf{h}_k(\mathbf{x}_k, \mathbf{n}_k)$,

$$\log P(\mathbf{z}_k|\mathbf{x}_k) = -0.5(\mathbf{z}_k - H\mathbf{x}_k)^T R^{-1}(\mathbf{z}_k - H\mathbf{x}_k) - 0.5|R| + C \tag{C.5}$$

where R is the covariance of the measurement model noise, $|R|$ is the determinant of R and C is a constant. If all the random variables were observed, the ML parameters could be solved by maximising the above equation. But the states are hidden so we use the expected values when we don't have access to the true values. The expected value of some variable $f(x)$ with respect to the posterior distribution of x is given by

$$\langle f(x) \rangle = \int_x f(x) P(x|z,\theta) dx \tag{C.6}$$

By taking derivatives to get a set of linear equations, the M-step for the measurement matrix is

$$H \leftarrow (\sum_k \mathbf{z}_k \langle \mathbf{x}_k \rangle^T)(\sum_k \langle \mathbf{x}_k \mathbf{x}_k^T \rangle)^{-1} \tag{C.7}$$

Similar steps are taken for all other parameters F, Q, R. The terms $\langle \mathbf{x}_k \rangle$,$\langle \mathbf{x}_k \mathbf{x}_k^T \rangle$ and $\langle \mathbf{x}_k \rangle$,$\langle \mathbf{x}_k \mathbf{x}_{k-1}^T \rangle$ are computed using Kalman Smoothing, which solves the problem of estimating the state at a given time step of a linear-Gaussian state-space model given the model parameters and a sequence of observations.

D

Algorithms for Hidden Markov Model decoding and evaluation

D.1 The Forwards Algorithm

The Forwards algorithm allows us to calculate the probability of the observation sequence $O = O_1, O_2, \ldots, O_T$ given a model $\Theta = (A, B, \Pi)$, i.e. $P(O|\Theta)$. This can be done most straightforwardly by computing the probability of every state sequence. If the state sequence is

$$Q = q_1, q_2, \ldots, q_T \tag{D.1}$$

the probability of the observation sequence, O, is (assuming statistical independence of the observations)

$$P(O|Q, \Theta) = \prod_{t=1}^{T} P(O_t|q_t, \Theta) \tag{D.2}$$

$$P(O|Q, \Theta) = b_{q_1}(O_1) b_{q_2}(O_2) \ldots b_{qT}(O_T) \tag{D.3}$$

The probability of this state sequence can be written as

$$P(Q|\Theta) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \ldots a_{q_{T-1} q_T} \tag{D.4}$$

Now the joint probability of O and Q is

$$P(O, Q|\Theta) = P(O|Q, \Theta) P(Q, \Theta) \tag{D.5}$$

The probability of O is found by marginalising i.e. summing over all possible state sequences q

$$P(O|\Theta) = \sum_{q_1, q_2, \ldots, q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \ldots a_{q_{T-1} q_T} b_{q_1}(O_T) \tag{D.6}$$

This calculation is interpreted as follows. At time $t = 1$ we are in state q_1 with probability π_{q_1}

and generate observation O_1. At $t = 2$, the transition is made to q_2 with probability $a_{q_1 q_2}$ and observation O_2 with probability $b_{q_2}(O_2)$. Which continues until the last transition.

This calculation is order $2TN^T$ since at every $t = 1, 2, \ldots, T$ there are N possible states which can be reaches (N^T possible state sequences) and for each state sequence there are $2T$ calculations required. This is therefore computationally unfeasible since for $N = 5$ and $T = 100$ there are 10^{72} computations.

A more efficient procedure exists. If we define the forward variable as

$$\alpha_t(i) = P(O_1, O_2, \ldots, O_t, q_t = S_i|\Theta) \tag{D.7}$$

which is the probability of the partial observation sequence, $O_1, O_2, \ldots, O_t$ until time t given the model Θ.

It is possible to solve for $\alpha_t(i)$ inductively

Initialisation

$$\alpha_1(i) = \pi_i b_i(O_1) \tag{D.8}$$

Induction

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] b_j(O_{t+1}), \tag{D.9}$$

Termination

$$P(O|\Theta) = \sum_{i=1}^{N} \alpha_T(i) \tag{D.10}$$

The calculation requires order N^2T calculations (i.e. $N(N+1)T(T-1)+N$ multiplications and $N(N-1)(T-1)$ additions) so for $N = 5, T = 100$ we need around 3000 calculations compared to 10^{72} for the exhaustive case.

D.2 The Viterbi Algorithm

The optimal state sequence associated with a set of observations is defined as the states q_t which are individually most likely, which is the optimality criterion which maximises the expected

number of correct individual states. To implement a solution to this we define two variables. The first is the *backwards* variable:

$$\beta_t(i) = P(O_{t+1}, O_{t+2}, \ldots, O_T | q_t = S_i, \Theta) \tag{D.11}$$

The second is the probability of being in state S_i at time t given the observation sequence, O, and the model, Θ, which can be expressed in terms of the forward and backward variables.

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O|\Theta)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \tag{D.12}$$

since $\alpha_t(i)$ accounts for the partial observation sequence $O_1, O_2, \ldots, O_t$ and state S_i at t, while $\beta_t(i)$ accounts for the remainder of the observation sequence $O_{t+1}, O_{t+2}, \ldots, O_T$ given state S_i at t. The normalisation factor $P(O|\Theta) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i)$ makes $\gamma_t(i)$ a true probability such that

$$\sum_{i=1}^{N} \gamma_t(i) = 1 \tag{D.13}$$

using $\gamma_t(i)$ we can solve for the individually most likely state q_t at time t

$$q_t = \underset{1 \leq i \leq N}{\arg\max}[\gamma_t(i)] \tag{D.14}$$

This equation maximise the individually most likely state at any instant. But there can be problems with the *state sequence*. If, for example, the HMM has transitions which have $a_{ij} = 0$ for some i, j i.e. state transitions with zero probability the "optimal" (using this optimality measure) state sequence may not even be valid.

The optimality criterion can therefore be modified to to find the single best *path* through the state trellis. This is the Viterbi algorithm.

The highest probability along a path, at time t which accounts for the first t observations and

ends in state S_i is

$$\delta_t(i) = [\max_{q_1,q_2,\cdots}, q_{t-1}]P(q_1, q_2, \ldots, q_t = i, O_i, O_2|\Theta) \tag{D.15}$$

By induction we have

$$\delta_{t+1}(j) = [\max_i \delta_t(i)a_{ij}]b_j(O_{t+1}) \tag{D.16}$$

In order to find the state sequence we need to track the argument $\delta_t(i)$. This is done via an array $\psi_t(j)$. The algorithm can be stated as follows:

Initialisation

$$\delta_t(i) = \pi_i b_i(O_1) \tag{D.17}$$

$$\psi_1(i) = 0 \tag{D.18}$$

Recursion

$$\delta_t(j) = \max_{1\leq i\leq N}[\delta_{t-1}(i)a_{ij}]b_j(O_t) \tag{D.19}$$

$$\psi_j = \arg\max_{1\leq i\leq N}[\delta_{t-1}(i)a_{ij}] \tag{D.20}$$

Termination

$$p^* = \max_{1\leq i\leq N}[\delta_T(i)] \tag{D.21}$$

$$q_T^* = \arg\max_{1\leq i\leq N}[\delta_T(i) \tag{D.22}$$

Best state tracking

$$q_t^* = \psi_{t+1}(q_{t+1}^*), t = T-1, T-2, \ldots, 1 \tag{D.23}$$

E

Optic flow computation

The Lucas-Kanade algorithm [89, 90] is a weighted least-squares fit of local first-order constraints to a constant model for velocity $\mathbf{v}$ in each small neighbourhood Ω by minimising

$$\sum_{\mathbf{x}\in\Omega} W_2(\mathbf{x})[\nabla I(\mathbf{I},t)\cdot\mathbf{v}+I_t(\mathbf{x},t)]^2 \tag{E.1}$$

where $W(\mathbf{x})$ is a window function which assigns higher weights to the centre of the neighbourhood. The solution to this is

$$A^TW^2A\mathbf{v}=A^TW^2\mathbf{b} \tag{E.2}$$

where, for n points $\mathbf{x}_i\in\Omega$ at time t,

$$\begin{aligned} A &= [\nabla I(\mathbf{x}_1),\ldots,\nabla I(\mathbf{x}_n)]^T \\ W &= \mathrm{diag}[W(\mathbf{x}_1),\ldots,W(\mathbf{x}_n)] \\ \mathbf{b} &= -[I_t(\mathbf{x}_1),\ldots,I_t(\mathbf{x}_n]^T \end{aligned} \tag{E.3}$$

The solution to E.2 is

$$\mathbf{v}=[A^TW^2A]^{-1}A^TW^2\mathbf{b} \tag{E.4}$$

This is solved in closed form when A^TW^2A is non-singular, since it is a 2×2 matrix:

$$A^TW^2A=\begin{bmatrix}\sum W^2(\mathbf{x})I_x^2(\mathbf{x}) & \sum W^2(\mathbf{x}I_x)(\mathbf{x})I_y(\mathbf{x})\\ \sum W^2(\mathbf{x}I_x)(\mathbf{x})I_y(\mathbf{x}) & \sum W^2(\mathbf{x})I_y^2(\mathbf{x})\end{bmatrix} \tag{E.5}$$

(All sums are over the points in the neighbourhood Ω.)

Bibliography

[1] E.L. Andrade, R.B. Fisher *Simulation of Crowd Problems for Computer Vision* First International Workshop on Crowd Simulation (V-CROWDS '05), Lausanne, Nov 2005.

[2] E.L. Andrade, S. Blunsden, R.B. Fisher *Characterisation of Optical Flow Anomalies in Crowd* IEE Int. Symp. on Imaging for Crime Detection and Prevention (ICDP 2005), London, pp 73-78, 7-8 June 2005.

[3] H. Attias *A variational Bayesian framework for graphical models*, Advances in Neural Information Processing Systems 12 (ed. T. Leen et al.), 2000, MIT Press, Cambridge, MA, USA.

[4] Z. Bar-Joseph, R. El-Yaniv, D. Lischinski, M. Werman *Texture Mixing and Texture Movie Synthesis Using Statistical Learning* IEEE Trans. Vis. Comput. Graph. 7(2): 120-135 (2001).

[5] J.L. Barron, D.J. Fleet, S.S. Beauchemin *Performance of Optical Flow Techniques* International Journal of Computer Vision 12:1 pp43-77, 1994.

[6] L. E. Baum and J. A. Egon *An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology* Bull. Amer. Meterorol. Soc., vol 73, pp. 360–363, 1967.

[7] L. E. Baum and G. R. Sell, *Growth functions for transformations on manifolds* Pac. J. Math., vol. 27, no. 2, pp. 211–227, 1968.

[8] J. S. Beis and D. G. Lowe *Shape indexing using approximate nearest-neighbour search in high-dimensional space* IEEE Conf. on Computer Vision and Pattern Recognition, San Juan, PR, June 1997.

[9] P.N. Belhumeur, J. Hespanha, and D. Kriegman *Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear rojection* IEEE Trans. PAMI, Special Issue on Face Recognition, 19(7), 11-20 (1997).

[10] C. Bibby and I. Reid *Visual Tracking at Sea* International Conference on Robotics and Applications, Barcelona, 2005

[11] C.M. Bishop *Variational principle component analysis*, Proc. 9th ICANN, 2000.

[12] J. Black, D. Makris, T.J. Ellis *Validation of Blind Region Learning and Tracking* Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation, Beijing, China, October, 2005.

[13] A. Blake and M. Isard *Active Contours*, 2nd ed., Springer-Verlag, London, 2000.

[14] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri *Actions as Space-Time Shapes* IEEE International Conference on Computer Vision (ICCV), Beijing, October 2005.

[15] V. Blanz, T. Vetter *Face Recognition Based on Fitting a 3D Morphable Model* IEEE Trans. Pattern Anal. Mach. Intell. 25(9): 1063-1074 (2003).

[16] A.F. Bobick *Computers seeing action*, Proc. British Machine Vision Conf., 1, 13-22, 1996.

[17] O. Boiman and M. Irani *Detecting Irregularities in Images and in Video* IEEE International Conference on Computer Vision (ICCV), Beijing, October 2005.

[18] M.Brand, L.Birnbaum, P.Cooper *Sensible scenes: visual understanding of complex structures through causal analysis*, Proceedings, National Conference on Artificial Intelligence, Washington D.C., 1993.

[19] M. Brand *Coupled hidden Markov models for modeling interacting processes* MIT Media Lab Perceptual Computing, Learning and Common Sense Techincal Report 405 (Revised), June 1997.

[20] M. Brand *Pattern discovery via entropy minimization* Proc. Uncertainty 1999 (AI and Statistics), 1999.

[21] M. Brand *Structure discovery in conditional probability models via an entropic prior and parameter extinction*, Neural Computation, 11, 5, 1155-1182, 1999.

[22] M. Brand and V. Kettnaker *Discovery and Segmentation of Actions in Video* IEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 8, August 2000.

[23] M.E. Bratman *Intention, Plans, and Practical Reason* CSLI Publications, Stanford University, ISBN (Paperback): 1575861925, 1988.

[24] W.L. Buntine *Operations for learning with graphical models*, Journal of Artificial Intelligence Research, 1994.

[25] H. Buxton and S. Gong *Advanced visual surveillance using Bayesian networks*, Proc. Int. Conf. Compter Vision, 1995.

[26] H. Buxton *Learning and Understanding Dynamic Scene Activity* ECCV Generative Model Based Vision Workshop, Copenhagen, Denmark, 2002.

[27] D. Chai and K. N. Ngan *Locating facial region of a head-and-shoulders color image* Third IEEE International Conference on Automatic Face and Gesture Recognitions, Nara, Japan, pp. 124-129, April 1998.

[28] T. Cham and R. Cipolla *Automated B-spline curve representation incorporating MDL and error-minimising control point insertion strategies*, IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 21 (1999), number 1, pp 49-53.

[29] H. Choi and B.J. Scholl *Effects of grouping and attention on the perception of causality* Perception & Psychophysics, 66(6), 926 - 942, 2004.

[30] R.T. Collins *Mean-shift Blob Tracking through Scale Space* IEEE Computer Vision and Pattern Recognition, Madison, WI, June 2003.

[31] D. Comaniciu and P. Meer *Mean Shift Analysis and Applications* Proceedings of the International Conference on Computer Vision-Volume 2, p.1197, September 20-25, 1999.

[32] L.Cooper and M.Munger *Extrapolating and remembering positions along cognitive trajectories: Uses and limitations of analogies to physical motion*, in [162].

[33] P.R.Cooper, M.A.Brand, A knowledge framework for seeing and learning, Visual Learning, Volume 2: Symbolic Visual Learning, Katsushi Ikeuchi and Manuela Veloso, eds. Oxford University Press, 1995.

[34] T. F. Cootes and K. Walker and C. J. Taylor *View-Based Active Appearance Models* FG '00: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000.

[35] R. Cucchiara, M. Piccardi and P. Mello *Image analysis and rule-based reasoning for a traffic monitoring system* IEEE Transactions on Intelligent Transportation Systems, June 2000 1(2) pp. 119-130.

[36] R. Cutler and L. Davis *Robust periodic motion and motion symmetry detection* IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), June 2000, Hilton Head Island, South Carolina.

[37] F. Cuzzolin, A. Bissacco, R. Frezza and S.Soatto *Towards unsupervised detection of actions in clutter*, Technical report, UCLA CSD-200033, December 2000, Dept. Computer Science, UCLA, Los Angeles, CA 90095.

[38] J.S. De Bonet *Multiresolution sampling procedure for analysis and synthesis of texture images* SIGGRAPH 1997: 361-368.

[39] H. Dee and D. Hogg *Detecting Inexplicable Behaviour* Proceedings of the British Machine Vision Conference, 2004.

[40] J. Deutscher, B. North, B. Bascle and A. Blake *Tracking through singularites and discontinuities by random sampling*, Proc. Int. Conf. Computer Vision, 1999.

[41] J. Driver, P. McLeod and Z. Dienes *Motion coherence and conjunction search: Implications for guided search theory*, Perception and Psychophysics 51, 1, 79-85, 1992.

[42] Duda and Hart *Pattern classification and scene analysis*, J. Wiley pub., 1973.

[43] N.Elian, R.McCarty and B.Brewer, (Eds) *Spatial Representations: Problems in Philosophy and Psychology*, Blackwell, 1993.

[44] A.A. Efros, A. Berg, G. Mori and J. Malik *Recognising Action at a Distance* Proceedings of the International Conference on Computer Vision, Nice, France, July 2003.

[45] I.A. Essa and A.O. Pentland *Facial expression recognition using a dynamical model and motion energy*, Proc. 5th Int. Conf. Computer Vision, pp 360-367, 1995.

[46] M. Everingham and A. Zisserman *Identifying individuals in video by combining generative and discriminative head models* Proceedings of the International Conference on Computer Vision, Beijing, China, October 17-20, 2005.

[47] J.H. Fernyhough, A.G. Cohn and D.C. Hogg *Building qualitative event models automatically from visual input*, Proc. Int. Conf. Compter Vision, 350-355, 1998.

[48] D.J. Fleet and A.D. Jepson *Computation of component image velocity from local phase information* Intenational Journal of Computer Vision, 5, pp 77-104.

[49] D.J Fleet *Measurement of Image Velocity* Kluwer Academic Publishers, Norwell.

[50] J. Forbes, T. Huang, K. Kanazawa, S.J. Russell *The BATMobile: Towards a Bayesian Automated Taxi* International Joint Conference on Artificial Intelligence, 1995, pp. 1878-1885.

[51] G. Foresti, L. Snidaro, P. Remagnino, T.J. Ellis Advanced Image and Video Processing in Active Video-Based Surveillance Systems IEEE Signal Processing Magazine, IEEE, 2005.

[52] A. Galata, N. Johnson, D. Hogg *Learning Behaviour Models of Human Activities* British Machine Vision Conference, 1999.

[53] A. Galata, A. Cohn, D. Magee, D. Hogg *Learning temporal and qualitative spatial components of an interaction model*, Proc. Vision and Modelling of Dynamic Scenes, in conjunction with The European Conf. Computer Vision 2002.

[54] A.H. Gee and R. Cipolla. *Determining the gaze of faces in images.* Image and Vision Computing, 12(10):639-647, December 1994.

[55] M. Georgeff, B. Pell, M. Pollack, M. Tambe, M. Wooldridge *The Belief-Desire-Intention Model of Agency* Proceedings of the 5th International Workshop on Intelligent Agents V : Agent Theories, Architectures, and Languages (ATAL-98).

[56] R. Gerber, H. Nagel, H. Schreiber Deriving Textual Descriptions of Road Traffic Queues from Video Sequences European Conference on Artifical Intelligence 2002, pp. 736-740.

[57] R. Gerber and H.-H. Nagel *'Occurrence' Extraction from Image Sequences of Road Traffic Scenes* Cognitive Vision Workshop, 19-20 September 2002, Zurich, Switzerland.

[58] A. Gersho and R.M. Gray *Vector Quantisation and Signal Compression*, Kluwer Academic Press, 1991.

[59] Z. Ghahramani and M.I. Jordan *Factorial Hidden Markov Models* In Adances in Neural Information Processing Systems, vol.8, Cambridge, MA, 1996, MIT Press.

[60] Z. Ghahramani *Learning Dynamic Bayesian Networks* In C.L. Giles and M. Gori (eds.), Adaptive Processing of Sequences and Data Structures. Lecture Notes in Artificial Intelligence, 168-197. Berlin: Springer-Verlag.

[61] S. Gilles *Description and experiment of image matching using mutual information*, Technical Report, Robotics Research Group, Oxford University, 1996.

[62] R.C. Gonzalez and R.E. Woods *Digital Image Processing*, Addison-Wesley Publishing Company, 1993.

[63] W.E.L. Grimson, C. Stauffer, R. Romano, and L. Lee. *Using Adaptive Tracking to Classify and Monitor Activities in a Site* Computer Vision and Pattern Recognition, June 23-25, 1998, Santa Barbara, CA, USA.

[64] T. R. Gruber *A translation approach to portable ontologies* Knowledge Acquisition, 5(2):199-220, 1993.

[65] R. Hartley and A. Zisserman *Mulitple view geometry in computer vision* Cambridge University Press, 2nd Ed., 2003, ISBN 0521 54051 8.

[66] D.J. Heeger and J.R. Bergen *Pyramid-based texture analysis/synthesis* SIGGRAPH 1995: 229-238.

[67] F.Heider and M.Simmel, *An experimental study of apparent behaviour*, Am. J. Psychol. 57, 243-249, 1944.

[68] Ho and Lee *A Bayesian approach to Problems in Stochastic Estimation and Control* IEEE Transactions on Automatic Control, pages 333– 339, October 1964.

[69] Douglas Hofstadter, *Gdel, Escher, Bach: an Eternal Golden Braid* Basic Books, 1979, ISBN 0465026567.

[70] B.K.P Horn *Robot Vision* MIT Press, Cambridge, MA, USA, 1986.

[71] R.J. Howarth *Interpreting a Dynamic and Uncertain World: Task-Based Control* Artifial Intelligence 100(1-2): 5-86 (1998).

[72] K. Hidai et al. *Robust Face Detection against Brightness Fluctuation and Size Variation* International Conference on Intelligent Robots and Systems, vol. 2 pp. 1379-1384, Japan, October 2000.

[73] K. Ikeuchi and T. Suehiro *Towards an assembly plan from observation pat i: Task recognition with poyhedral objects* IEEE Trans. Robotics and Automation, 10(3):368-385, 1994.

[74] *MAIA : Autonomous intelligent machine* http://www.inria.fr/recherche/equipes/maia.en.html.

[75] C. Jacobs, A. Finkelstein, D. Salesin *Fast Multiresolution Image Querying* Computer Graphics, Annual Conference Series (Siggraph'95 Proceedings), pp. 277-286, 1995.

[76] T.S. Jebara and A. Pentland *Parametrized Structure from Motion for 3D Adaptive Feedback Tracking of Faces* Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, pp. 144-150.

[77] G. Johannson *Visual perception of biological motion and a model for its analysis*, Perception and Psychophysics 14, 2, 201-211, 1973.

[78] N. Johnson and D. Hogg. *Learning the Distribution of Object Trajectories for Event Recognition* Proc. British Machine Vision Conference, vol. 2, pp. 583-592, September 1995.

[79] M.I. Jordan, Z. Ghahramani and L.K. Saul *Hidden Markov Decision Trees* In Adances in Neural Information Processing Systems, vol.8, Cambridge, MA, 1996, MIT Press.

[80] J. Y. Kaminski, M. Teicher, D. Knaan and A. Shavit *Three-Dimensional Face Orientation and Gaze Detection from a Single Image*, CoRR, cs.CV/0408012, 2004.

[81] K. Kanazawa, D. Koller, S.J. Russell *Stochastic Simulation Algorithms for Dynamic Probabilistic Networks* UAI '95: Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence, August 18-20, 1995, Montreal, Quebec, Canada.

[82] M. Kass et al. *Snakes: Active contour models*, Proc. Int. Conf. Comp. Vision, 1987, pp 259-268.

[83] B. Katz, S. Felshin, D. Yuret, A. Ibrahim, J. Lin, G. Marton, A. McFarland, and B. Temelkuran *Omnibase: Uniform Access to Heterogeneous Data for Question Answering* Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems, June 2002, Stockholm, Sweden.

[84] B. Katz, J. Lin, C. Stauffer, E. Grimson *Answering Questions about Moving Objects in Surveillance Videos* AAAI Spring Symposium on New Directions in Question Answering, March 2003.

[85] D. Koller, J. Weber, T. Huang, J. Malik, G. Ogasawara, B. Rao, S. Russell *Towards Robust Automatic Traffic Scene Analysis* In Real-Time In Proc. of the 12th International Conference on Pattern Recognition (ICPR-94) pp. 126-131, Jerusalem, Israel, October 9-13, 1994.

[86] B.Kuipers, Qualitative Reasoning, 1994 MIT Press, Cambridge, Massachusetts, USA.

[87] P.S. Laplace *A Philosophical Essay on Probabilities*, unabridged and unaltered reprint of Truscott and Emory translation, Dover Publications, Inc., New York, 1951, original publication date 1814.

[88] Y. Leclerc *Constructing simple stable descriptions for image partitioning*, Int. Journ. Computer Vision. 3 73-102 (1989).

[89] B.D. Lucas and T. Kanade *An Iterative Image Registration Technique with Application to Stereo Vision* DARPA Image Understanding Workshop, 1981.

[90] B.D. Lucas and T. Kanade *Generalized Image Matching by the Method of Differences* Ph.D, thesis, Department of Computer Science, Carnegie-Mellon Univeristy, 1984.

[91] D. Makris, T.J. Ellis *Learning Semantic Scene Models from Observing Activity in Visual Surveillanc* IEEE Transactions on Systems Man and Cybernetics - Part B 35(3) June, pp. 397-408. ISBN/ISSN 1083-4419, 2005.

[92] D. Makris and T.Ellis *Spatial and Probabilistic Modelling of Pedestrian Behaviour* British Machine Vision Conference 2002, vol. 2, pp. 557-566, Cardiff, UK, September 2-5, 2002.

[93] R. Mann, A. Jepson, J.M. Siskind *Computational Perception of Scene Dynamics* Proc. European Conference on Computer Vision, Cambridge, UK, 1996.

[94] Y. Matsumoto and A. Zelinsky *An Algorithm for Real-time Stereo Vision Implementation of Head Pose and Gaze Direction Measurement* Proceedings of IEEE Fourth International Conference on Face and Gesture Recognition, pp. 499-505, 2000.

[95] S.J. Maybank and P.F. Sturm *Minimum description length and the inference of scene structure from images*, In IEE Colloquium on Applied Statistical Pattern Recognition, pp. 9–16, 1999.

[96] P. McLeod, J. Driver and J. Crisp *Visual search for a conjunction of movement and form is parallel*, Nature 332, 154-155, 1988.

[97] J. McNames *A Fast Nearest-Neighbor Algorithm Based on a Principal Axis Search Tree* IEEE Pattern Analysis and Machine Intelligence, vol. 23, September 2001, pp. 964-976 ISSN:0162-8828.

[98] A. Michotte, *The perception of causality*, Basic Books, 1946. (English transl.) Methuen, Andover, MA, 1963.

[99] A.Michotte, *The emotions regarded as functional connections* Feelings and emotions: The Mooseheart symposium, pp114-125, McGraw-Hill, 1991.

[100] C. Mikolajczyk, C. Schmid and A. Zisserman *Human detection based on a probabilistic assembly of robust part detectors* Proc. European Conf. on Computer Vision, Prague, Czech Republic, May 2004.

[101] V. Morellas, I. Pavlidis, P. Tsiamyrtzis *DETER: Detection of Events for Threat Evaluation and Recognition* Machine Vision and Applications, 15(1):29-46, October 2003.

[102] A. Nairac, T. Corbett-Clark, R. Ripley, N. Townsend and L. Tarassenko. *Choosing an appropriate model for novelty detection* Proc 5th IEE Int. Conf on Artificial Neural Networks, Cambridge, 117-122 (1997).

[103] Nairac A, Townsend N, Carr R, King S, Cowley P and Tarassenko L. *A system for the analysis of Jet Engine Vibration Data* Integrated Computer-Aided Engineering, 6, 53-65 (1999).

[104] K. Nakayama, G. Silverman *Serial and parallel processing of visual feature conjunctions*, Nature 320, 264-265, 1986.

[105] S. A. Nene and S. K. Nayar *A Simple Algorithm for Nearest Neighbor Search in High Dimensions* IEEE Transactions on Pattern Analysis and Machine Intelligence vol.19, September 1997, p. 989-1003.

[106] R. Osadchy, M. Miller and Y. Lecun *Synergenistic face detection and pose estimation with energy-based model* Proc. NIPS 2005.

[107] N.M. Oliver, B. Rosario and A.P. Pentland *A Bayesian computer vision system for modelling human interactions*, IEE Trans. Pattern Analysis and Machine Intelligence, vol.22, No.8, August 2000.

[108] N. Oliver *Towards perceptual intelligence: statistical modelling of human individual and interactive behaviours*, PhD thesis, MIT, Media lab, Cambridge, Mass, 2000.

[109] D. Pang, M.D. and V. Li, M.D. *Atlantoaxial Rotatory Fixation: Part 1-Biomechanics OF Normal Rotation at the Atlantoaxial Joint in Children.* Neurosurgery. 55(3):614-626, September 2004.

[110] J. Pearl *Bayesian Networks: A model of self-activated memory for evidential reasoning* Proc. Cognitive Science Society, pp. 329-34, Greenwich, CT.

[111] J. Pearl *Causality. Models, Reasoning and Inference* Cambridge University Press, 2000, ISBN 0 521 77362 8.

[112] W.D. Penny and S.J. Roberts *Variational Bayes for 1-dimensional mixture models*, Technical report PARG-2000-01, Dept.Eng.Sci., Oxford University, 2000.

[113] R. Penrose *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics* Oxford Univ. Press, 1989.

[114] A. Pentland, B. Moghaddam and T. Starner *View-Based and Modular Eigenspaces for Face Recognition* Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'94).

[115] A.Perez, M.L. Cordoba, A. Garcia, R. Mendez, M.L. Munoz, J.L. Pedraza, F. Sanchez *A Precise Eye-Gaze Detection and Tracking System* Proceedings of the 11th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision 2003.

[116] S.L. Phung, A. Bouzerdoum, and D. Chai *Skin segmentation using color and edge information* Proc. Int. Symposium on Signal Processing and its Applications, 1-4 July 2003, Paris, France.

[117] R. W. Picard *Digital Libraries: Meeting Place for High-level and Low-level Vision* Invited paper, Proc. Asian Conference on Computer Vision, Singapore, Dec 1995.

[118] F. Porikli and T. Haga *Event Detection by Eigenvector Decomposition Using Object and Frame Features* Computer Vision and Pattern Recognition, Washington D.C., USA, June 2004.

[119] J.D. Prothero and H.G. Hoffman *Widening the Field-of-View Increases the Sense of Presence in Immersive Virtual Environments* Technical Report TR-95-2, Human Interface Technology Laboratory, University of Washington.

[120] L.R. Rabiner *A tutorial on Hidden Markov Models and selected applications in speech recognition*, Proc IEEE, 77, 2, 257-285, 1989.

[121] C.W.Reynolds *Flocks, Herds and schools: A distributed behavioural model*, Computer graphics 21, 4 (August), 25-34, 1987.

[122] J.R. Renno, D. Makris, T.J. Ellis, G.A. Jones *Application and Evaluation of Colour Constancy in Visual Surveillance* Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation, VS-PETS Beijing, China, October, 2005.

[123] I. Rezek and S. Roberts. *Learning Ensemble Hidden Markov Models for Biosignal Analysis* 14th International Conference on Digital Signal Processing, Santorini, Greece, 2002.

[124] M. Rigolli, *D.Phil. thesis*, Department of Engineering Science, University of Oxford, 2006.

[125] M. Rigolli and M. Brady *Towards a Behavioural Traffic Monitoring System* Autonomous Agents and Multi-agent Systems, Utrecht, Netherlands, July 2005.

[126] M. Rigolli, Q. Williams, M.J. Gooding and M. Brady *Driver Behavioural Classification from Trajectory Data* IEEE Conference on Intelligent Systems, Vienna, Austria, September 2005.

[127] J.J. Rissanen *Modelling by the shortest data description*, Automatica-J.IFAC 14 (1978), pp 465-471.

[128] S.J. Roberts, D. Husmeier, I. Rezek and W. Penny *Bayesian approaches to Gaussian mixture modelling*, Pattern Analysis and Machine Intelligence, 20 (11), 1133-1142, 1998.

[129] P. Robertson, D.Phil. thesis, University of Oxford, Dept. of Engineering Science, 2001.

[130] N.M. Robertson, I.D. Reid and J.M. Brady *What are you looking at? Gaze recognition in medium-scale images* Human Activity Modelling and Recognition (HAREM), British Machine Vision Conference (BMVC), Oxford, UK, September 2005.

[131] N.M. Robertson and I.D. Reid *Behaviour understanding in video: a combined method* Proceedings of the International Conference on Computer Vision (ICCV), October 2005, Beijing, China.

[132] N.M. Robertson and I.D. Reid *Estimating Gaze Direction from Low-Resolution Faces in Video* Proceedings of the 9th European Conference on Computer Vision (ECCV), Graz, Austria, May 2006.

[133] N.M. Robertson and I.D. Reid *Human activity recognition in video using a combination of parametric and non-parametric techniques* Journ. Computer Vision and Image Understanding (CVIU), Special Issue on Modeling People: Shape, Appearance, Movement and Behaviour, 2006.

[134] N.M. Robertson, I.D. Reid and J.M. Brady *Behaviour Recognition and Explanation for Video Surveillance* Imaging for Crime Detection and Prevention (ICDP), June 2006, London

[135] S. Romdhani *Face Image Analysis using a Multiple Feature Fitting Strategy* PhD Thesis, Unversity of Basel, January 2005

[136] S. Russel and P. Norvig *Artificial intelligence, a modern approach* Prentice-Hall, 1995.

[137] L.K. Saul and M.I. Jordan *Boltzmann Chains and Hidden Markov Models* In Adances in Neural Information Processing Systems, vol.8, Cambridge, MA, 1996, MIT Press.

[138] A.Schlottman, L.Surian, *Do 9-month-olds perceive causation at a distance?*, Perception, 28, 1105-1114, 1999.

[139] H. Schneiderman and T. Kanade *A statistical method for 3D object detection applied to faces and cars* Computer Vision and Pattern Recognition, 2000.

[140] B.J.Scholl and P.D.Tremoulet, *Perceptual causality and animacy*, Trends in Cognitive sciences, vol.4, No.8 Aug. 2000, Elsevier Science.

[141] B.J. Scholl and K. Nakayama *Causal capture: Contextual effects on the perception of collision events* Psychological Science, 13(6), 493 - 498, 2002.

[142] B.J. Scholl *Innateness and (Bayesian) visual perception* In P. Carruthers, S. Laurence and S. Stich (Eds.), The innate mind: Structure and contents (pp. 34 - 52), Oxford University Press, 2005.

[143] R.Shor, *Effect of preinformation upon human characteristics attributed to animated geometric figures*, J. Abnorm. Soc. Psychol. 54, 124-126.

[144] H. Sidenbladh M. Black, L. Sigal. *Implicit Probabilistic Models of Human Motion for Synthesis and Tracking* European Conference on Computer Vision, Copenhagen, Denmark, June 2002.

[145] N.T. Siebel *Fusion of Multiple Tracking Algorithms for Robust People Tracking* European Conference on Computer Vision, May 2002, Copenhagen, Denmark.

[146] W. Siler and J.J. Buckley *Fuzzy Expert Systems and Fuzzy Reasoning* ISBN: 0-471-38859-9, January 2005.

[147] J.M. Siskind *Naive Physics, Event Perception, Lexical Semantics and Language Acquisition* Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, January 1992.

[148] C. Stauffer, E. Grimson *Learning Patterns of Activity Using Real-Time Tracking* Pattern Analysis and Machine Intelligence, 22(8):747-757, 2000.

[149] C. Stauffer *Estimating Tracking Sources and Sinks* Proceedings of the Second IEEE Workshop on Event Mining, July 17, 2003.

[150] R.S. Sutton and A.G. Barto *Reinforcement Learning: An Introduction* MIT Press, Cambridge, MA, 1998 A Bradford Book.

[151] M. Tambe, J. Adibi,Y. Alonaizon, A. Erdem, G. Kaminka, S. Marsella, and I. Muslea, *Building agent teams using an explicit teamwork model and learning* Artificial Intelligence, volume 110, pages 215-240, 1999.

[152] L. Tarassenko, A. Nairac, N. Townsend, I. Buston and P. Cowley. *Novelty Detection for the Identification of Abnormalities* International Journal of Systems Science, 11, 1427-1439 (2000).

[153] C.P. Town *Ontology-driven Bayesian Networks for Dynamic Scene Understanding* Proc. International Workshop on Detection and Recognition of Events in Video (at CVPR04), 2004.

[154] M. Turk and A. Pentland *Face recognition using eigenfaces* Proc. IEEE Conference on Computer Vision and Pattern Recognition, Maui, Hawaii, 1991.

[155] S.Ullman, *The interpretation of structure from motion*, Proc. R. Soc. London Ser. B 203, 405-426, 1979.

[156] A. Verri and T. Poggio *Against quantitative optical flow* Proc. IEEE International Conference on Computer Vision, London, 1987, pp. 171-180.

[157] P. Viola and W.M. Wells *Alignment by Maximisation of Mutual Information*, Int. Journal Computer Vision, 24(2) pp 137-154, 1997.

[158] P. Viola, M. Jones, D. Snow *Detecting Pedestrians using Patterns of Motion and Appearance* Proceedings of the International Conference on Computer Vision, Nice, France, July 2003.

[159] P.A. Viola and M.J. Jones *Robust Real-Time Face Detection* International Journal of Computer Vision,57(2),2004 pp. 137-154.

[160] P. Vitanyi and M. Li *Minimum description length induction*, Bayesianism and Kolmogorov Complexity, 1998.

[161] C.S. Wallace and P.R. Freeman *Estimation and inference by compact coding*, J. Royal Stat. Soc., Series B, 49 (1987) pp 240-251.

[162] D.S.Weld, J.de Kleer, (Eds), *Qualitiative Reasoning about Physical Systems*, Pub. Morgan Kaufmann, Palo Alto, CA, USA, 1990.

[163] B.B. Werger *Cooperation without deliberation: A minimal behavior-based approach to multi-robot teams* Artificial Intelligence, 110:293–320, 1999.

[164] Li-Yi Wei and Marc Levoy *Fast Texture Synthesis using Tree-structured Vector Quantization* In Proceedings of SIGGRAPH 2000.

[165] P.A.White, *Causal processing: origins and development*, Psychol. Bull. 104, 36-52, 1988.

[166] T. Xiang and S. Gong *Video behaviour profiling and abnormality detection without manual labelling* In Proc. International Conference on Computer Vision (ICCV), Beijing, China, October 2005.

[167] T. Xiang and S. Gong *Visual learning given sparse data of unknown complexity* In Proc. International Conference on Computer Vision (ICCV), Beijing, China, October 2005.

[168] L. Zelnik-Manor and M. Irani *Event-Based Video Analysis* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), December 2001.

[169] H. Zhong, J. Shi and M. Visontai *Detecting Unusual Activity in Video* Computer Vision and Pattern Recognition, Washington D.C., USA, June 2004.