FROM VISUAL SALIENCY TO VIDEO BEHAVIOUR UNDERSTANDING

Hayley Shi Wen Hung

Submitted to the University of London in partial fulfillment of the requirements for the degree of Doctor of Philosophy

QUEEN MARY, UNIVERSITY OF LONDON

2007

Abstract

In a world of ever increasing amounts of video data, we are forced to abandon traditional methods of scene interpretation by fully manual means. Under such circumstances, some form of automation is highly desirable but this can be a very open ended issue with high complexity. Dealing with such large amounts of data is a non-trivial task that requires efficient selective extraction of parts of a scene which have the potential to develop a higher semantic meaning, alone, or in combination with others. In particular, the types of video data that are in need of automated analysis tend to be outdoor scenes with high levels of activity generated from either foreground or background. Such dynamic scenes add considerable complexity to the problem since we cannot rely on motion energy alone to detect regions of interest. Furthermore, the behaviour of these regions of motion can differ greatly, while still being highly dependent, both spatially and temporally on the movement of other objects within the scene. Modelling these dependencies, whilst eliminating as much redundancy from the feature extraction process as possible are the challenges addressed by this thesis.

In the first half, finding the right mechanism to extract and represent meaningful features from dynamic scenes with no prior knowledge is investigated. Meaningful or salient information is treated as the parts of a scene that stand out or seem unusual or interesting to us. The novelty of the work is that it is able to select salient scales in both space and time in which a particular spatio-temporal volume is considered interesting relative to the rest of the scene. By quantifying the temporal saliency values of regions of motion, it is possible to consider their importance in terms of both the long and short-term. Variations in entropy over spatio-temporal scales are used to select a context dependent measure of the local scene dynamics. A method of quantifying temporal saliency is devised based on the variation of the entropy of the intensity distribution in a spatio-temporal volume over incraeasing scales. Entropy is used over traditional filter methods since the stability or predictability of the intensity distribution over scales of a local spatio-temporal region can be defined more robustly relative to the context of its neighbourhood, even for regions exhibiting high intensity variation due to being extremely textured. Results show that it is possible to extract both locally salient features as well as globally salient temporal features from contrasting scenerios.

In the second part of the thesis, focus will shift towards binding these spatio-temporally salient features together so that some semantic meaning can be inferred from their interaction. Interaction in this sense, refers to any form of temporally correlated behaviour between any salient regions of motion in a scene. Feature binding as a mechanism for interactive behaviour understanding is particularly important if we consider that regions of interest may not be treated as particularly significant individually, but represent much more semantically when considered in combination. Temporally correlated behaviour is identified and classified using accumulated co-occurrences of salient features at two levels. Firstly, co-occurrences are accumulated for spatio-temporally proximate salient features to form a local representation. Then, at the next level, the co-occurrence of these locally spatio-temporally bound features are accumulated again in order to discover unusual behaviour in the scene. The novelty of

this work is that there are no assumptions made about whether interacting regions should be spatially proximate. Furthermore, no prior knowledge of the scene topology is used. Results show that it is possible to detect unusual interactions between regions of motion, which can visually infer higher levels of semantics.

In the final part of the thesis, a more specific investigation of human behaviour is addressed through classification and detection of interactions between 2 human subjects. Here, further modifications are made to the feature extraction process in order to quantify the spatio-temporal saliency of a region of motion. These features are then grouped to find the people in the scene. Then, a loose pose distribution model is extracted for each person for finding salient correlations between poses of two interacting people using canonical correlation analysis. These canonical factors can be formed into trajectories and used for classification. Levenshtein distance is then used to categorise the features. The novelty of the work is that the interactions do not have to be spatially connected or proximate for them to be recognised. Furthermore, the data used is outdoors and cluttered with non-stationary background. Results show that co-occurrence techniques have the potential to provide a more generalised, compact, and meaningful representation of dynamic interactive scene behaviour.

Acknowledgements

My interest for research would have not been fueled so well if it not for the helpful colleagues with whom I worked during this PhD. Many interesting discussions have helped to progress the thoughts and findings within the pages of this thesis. I would like to thank them for their encouragement and support. In addition I would like to thank the many who supported me pastorally and helped me to maintain a relatively balanced life despite the pressures of doing a PhD. I would also like to thank those who enabled the final days of thesis process to run smoothly. I list them here in alphabetical order:

Samer Abdallah, Lourdes Agapito, Bushra Akhtar, Andrew Anderson, Keith Anderson, Melanie Aurnhammer, Sileye Ba, Simon Bance, Gianluigi Bellin, Matt Bernstein, Keiran Betteley, Alessio del Bue, Paul Brossier, Ewan Campbell, Helen Cassidy, Keith Clarke, Bobby Demers, Jim Divers, Michela Farenza, Renata Favalli, Katayoun Ferrahi, Christopher Frauenberger, Pierre le Fur, Jose Galan, Andrew Graves, Neill Hadlow, Chris Harte, David Hawes, Richard Howarth, Pat Healey, David Holoway, Dinesh Jayagopi, Hiren Joshi, Tim Kay, Tom King, Jia Kui, Mounia Lalmas, Joe Leach, Li Jun, Xavier Llado, Pasquale Malacaria, Robert Marchington, Peter McOwan, Christof Monz, Shahin Nabavian, Radu-Andrei Negoescu, Hugo Noda, Katy Noland, Sapna Nundloll, Samuel Pachoud, Marios Panayides, George Papatzanis, Leung Po, Pedro Quelhas, Julie Ringham, Edmund Robinson, David Russell, Paulo Santos, Shan Caifeng, Adam Sherwood, Fabrizio Smeraldi, Kevin Smith, Tassos Tombros, Chrystie Myketiak, Shahin Nabavian, Louise Nickerson, Jean-Baptiste Thiebaut, Milan Verma, Wang Yong, Simon Wells, Graham White, Sue White, Xiang Tao, and Lucasz Zalewski. I would also like to thank some very kind undergraduate tutees for volunteering their time to collect video date for some of my experiments. I would also like to thank them for the persepective they brought to my PhD when I was teaching them.

I would like to thank my examiners David Hogg and Maria Petrou for taking time to read my thesis carefully, their useful comments, suggestions and discussions.

I would also like to thank my family for supporting me throughout my endeavours.

And finally, I would like to thank my supervisor, Sean Gong for his energy, enthusiasm and encouragement.

This research was funded by the EPRSC, part-funded by QinetiQ Ltd and a travel grant was also contributed by RAEng.

Contents

1 Introduction

1.1 Automatic scene analysis

In the last decade, security cameras have become a common sight in the urban landscape. The ability to monitor a number of different places from one location has aided crime prevention. A vast number of security cameras are being installed in everything from public streets to train carriages, which has spawned a fundamental problem. On the London Underground alone, there are at least 9000 security cameras (McCahill and Norris, June 2002). Security staff can have as many as 60 cameras to watch at any one time so monitoring is an extremely difficult task, requiring considerable concentration for long periods of time (Donald, 1999). It is easy to imagine that manpower on its own is not enough to deal with the vast quantities of data that are being recorded. Automation is clearly the next step.

Automation is rather a vague term here. In an ideal world, we would prefer to automate surveillance as much as possible, whilst still preserving an individual's personal privacy. Furthermore, there are many different aspects of behaviour that we would like to automate. For example, unusual behaviour detection, criminal behaviour detection, person identification, object tracking, background modeling, to name just a few facets of this interesting problem. The work contained here cannot cover every application but from a computational perspective, there are many applications that still require distinctly similar stages of video, image and feature processing which will be exploited in the studies examined in the remainder of the thesis.

If we start from the very beginning of the automation process, we can begin to explore its problems, complexities and challenges. At the first stage, a video camera captures a quantised

representation of the real world which is already distorted from the original physical scene. The higher the resolution of the camera, the more photo-receptive elements are used per inch. Therefore, a more precise representation of the captured scene is possible. If we think of a video as a sequence of many images, where each image is made up of a grid of tens of thousands of pixels, it is easy to see that performing multiple operations on hundreds of thousands of pixels per second is a challenge in terms of computational complexity.

Sub-sampling on its own only serves to reduce computation while still having an inherent redundancy as much of a scene will contain information that is not interesting or useful to us[1]. Since we wish to extract meaning from the image or scene, a more specific means of selecting the important or significant features for further computation must be found. One method would be to find the Nyquist frequency for the local neighbourhood of each pixel of an image or video and through this process, it is very likely that contrasting regions would emerge with similar Nyquist frequencies. We might observe that the frequency value relates to the level of activity but also the context in which to observe the particular region. The ability to find regions of similarity allows us to determine topographical, structural as well as salient parts of a scene. However, grouping pixels is non-trivial especially if we do not know the number of regions and the groups do not form discrete sets.

1.2 Extracting meaningful activity

A more straight forward method of quantifying different levels of activity in image sequences would be to use motion detection methods. In most cases, we can assume that in video sequences, the areas of interest are most likely to be moving. However, in natural scenes, we soon find that the level of activity or changes in the signal do not necessarily equate to the meaningful or important information. Natural scenes contain much motion which may be considered as the background of the scene such as rippling water, moving foliage or reflections. So distinguishing between regions of interest in terms of stationary as well as non-stationary is less straight forward than just separating non-motion from motion. Despite this challenge,

[1] The word 'interesting' is used rather loosely here to describe information which appears distinctive. More specific ideas of what is interesting and useful will be described in more detail later

finding a method of grouping different spatial regions of the frame into significant and more semantically meaningful parts provides a more efficient foundation for further automated activity inference.

A question one might ask now is why we are trying to extract features without prior models. It may be argued that if we had prior models of the sorts of features we were interested in observing, such as people or cars, then we could use object detection to find the relevant regions of the image. This makes the fundamental assumption that we already know what we are looking for in the first place, which may be true in some cases but does not allow for the flexibility to identify much more elusive activity or behaviour that we may not have thought of modeling beforehand. Furthermore, objects in a scene may be partially occluded and finding robust models for object categorisation under lighting, view angle, as well as partial occlusion is impractical.

Assuming that it is possible to detect and extract meaningful features from a sequence, there are still some fundamental questions to address. We have already discussed the idea of extracting meaningful features even if we don't know what we are looking for. Therefore, more reliance is placed on trying to interpret the raw data and in particular the spatio-temporal dynamics of it. For an image, the parts which are likely to contain much information about the scene are areas with spatial change. For many images of natural scenes, there is continual spatial variation of pixel intensities. If we try to find more prominent types of variation, it is possible to identify turning points in the pixel intensities, which often represent the boundaries of regions of interest. The idea of using turning points is particularly significant since it describes a point at which some sequence of data becomes unpredictable. Using this idea, it is possible to find meaningful or salient features in both space and time. However, whether something is unpredictable or not is completely dependent upon its spatio-temporal context. That is, something which is salient over a five second time interval may not be so over a five minute, week or month time period. So finding the correct context or scale at which to observe a particular salient region requires an effective method of modeling its temporal change. Therefore, how temporal change is modeled is vital for finding discontinuities in feature space that correlate to contextual changes in the motion patterns.

Let us assume now that we are at a stage where it is possible to extract salient objects from a scene with sensible spatio-temporal context. The next stage is to identify a method to analyse the behaviour of these salient regions. The challenge is that image sequences of natural scenes are captured on a two-dimensional surface and thus it is prone to occlusion and restricted by view angle. Even if it was possible to make the assumption that we knew exactly what sorts of behaviours we wished to detect, finding accurate models would be difficult, given that the behaviours might occur at any location as well as viewing angle within the scene. To detect these behaviours accurately, it would be necessary to find exhaustive examples of a particular action, performed at multiple view angles to handle different degrees of occlusion in order to capture the full extent of the behaviour. For many natural scenes, however, such as a busy town high street, cluttered activity means that relying on finding continuous motion in order to identify salient homogeneous regions is clearly impractical.

It is likely that relying on tracking methods to find salient activity may not be the most sensible method of analysing naturally cluttered scene data which is compounded by the problem of camera placement. Surveillance cameras are placed for human operators. In open areas, the cameras tend to be placed in order to maximise the coverage area and therefore spatial resolution is compromised since objects are further away from the camera. In addition, camera placement can lead to unnecessary occlusion problems which might be avoided by using a different camera angle. Camera placement has become a serious consideration when presenting the motivations for the ideas in this thesis. If we are to use real camera sources rather than setting up different cameras for automated surveillance, tasks such as robust object tracking in crowded scenes can become an impossible task.

The other issue with camera placement is that despite the large numbers of cameras that are installed, it is still likely that some parts of a public space will not be covered. One would expect under such circumstances that it would be impossible to infer any information about what is happening in areas outside of the range of the cameras but activity which is off screen or occluded can still affect the visible part of a scene. Finding a method of using the visible parts of the scene to inform us of activity in invisible parts would be extremely powerful.

1.3 From feature binding to semantics

The problems highlighted so far are significant but even if these were optimised for automation, we are still faced with the problem of finding an effective way to model and understand natural scene activity. In many cases, repetition of a particular activity is not always reproducible in a spatial sense. That is, not all activities have significantly different spatial configurations that are discriminative. Furthermore, modeling *salient* activity is difficult since salient activities tend to occur less frequently. Therefore finding multiple examples for training is impractical if not impossible. Again, temporal context must be taken into account and can be illustrated with a simple example of a very busy scene of a market at 10am, as shown in Figure 1.1(a). At this time many people are walking around but if we were to revisit the market at 1am, as shown in Figure 1.1(b), clearly the idea of someone walking around will seem unusual since the market is empty, even though it would be considered normal at around 10 am. Context also plays a key role in how models of normal behaviour are constructed which makes making robust and reliable models even more difficult.

Another problem with modeling activities is that in many cases, organised crime is well informed of some deployed automated systems. Given the turnaround for developing and making a practically available system, it is very easy to see that criminals will always be one step ahead. Since technology cannot keep up, relying on trained models for finding criminal behaviour may be unrealistic for real practical applications. An example of a challenging activity to model is the three car-handle problem. This is extremely difficult to model since the objective of the criminal activity is to steal a car by trying different car handles on a street or car park. The activity can involve one or more people working in collaboration. Once an open car has been found, all assailants do not go to the car at once but will leave the car park and enter the car without drawing attention to themselves. In such circumstances it may be more practical to detect such activity by monitoring for temporally correlated behaviour. Finding groups of spatially separated but interacting people is non-trivial when spatial configuration is unconstrained and even temporal ordering is highly variable. In such cases, the main thing we can rely on is some form of temporal synchrony amongst salient

(a)

(b)

Figure 1.1: Two contrasting scenes from two markets. (a) is taken during the day and shows many people in the scene while (b) shows an empty market in the evening. The cases highlight that the temporal context plays an important role in deciding what is unusual or interesting in a scene.

events in time.

In summary the discussions above lead to the conclusion that one of the major issues within computer vision encompasses spatio-temporal binding. Whatever the applications are, finding suitable binding of the features or other sources of information is a challenging and interesting problem. I refer here to the binding as a process in the spatio-temporal domain but this would perhaps limit the depth and breadth of the problem. Defining spatial and temporal binding together is in itself neglecting the fact that binding in each domain may need to be done separately before they can be combined. On the other hand, binding spatial features may require temporal context and similarly, temporal binding requires spatial contexts and deciding in which order the processing needs to be done is also non-trivial.

1.4 The approach

1.4.1 What is saliency?

The philosophy of the work in this thesis is based on the idea of trying to extract as much information from a scene before top-down discriminative models need be applied. I have already described the pitfalls of relying on models, particularly in natural scenes. Therefore the philosophy of the work is to maximise knowledge extraction from video data through finding the underlying spatio-temporal dynamics of it.

The main theme of the work described in the following chapters involves using a saliency measure to discover contextually, what is significant in a spatial, temporal and/or spatio-temporal context. If a mechanism is available to quantify saliency, the information or data from a particular scene has been mapped to a feature domain in which some level of inference about the raw data can be extracted. Saliency is a particularly useful concept when considering bottom-up feature extraction since one must find what is significant in an image or sequence from the scene data alone. In such circumstances, the role of context becomes extremely important. That is to say that saliency can only be described as a relative measure of importance.

In the work described here, saliency is defined firstly in terms of spatial, then spatio-

temporal homogeneity. Therefore, regions that exhibit spatial or spatio-temporal homogeneity must inevitably end or meet a contour around which the pixel intensities differ more than is expected. Finding edges gives us spatial and/or temporal context in which to consider the region.

It is important at this point to make a minor digression to discuss the importance of edges in relation to image saliency. Early methods of feature extraction tended to concentrate on edge extraction methods for finding significant features within a scene. In simple cases, this is an effective method of finding spatial structure. Ultimately we are more interested in discovering the object itself from which the edges originate. Constructing such information from edges is particularly difficult for non-rigid objects. Since the natural world consists largely of deformable shapes, finding edges in order to discover spatially homogeneous objects seems an over exertion of computational resources. Rather, we may wish to separate foreground from background (stationary and non-stationary) before observing parts of a scene in more detail.

1.4.2 Binding features by co-occurrence

Finding an effective method of binding features is an essential part of interpreting either images or video. Combining features spatially, in a meaningful way, is already a difficult problem and one can already imagine that doing so temporally would be even more challenging. Furthermore, if we are interested in modeling activities, then very soon it becomes apparent that the same activities can occur over very different degrees of spatial separation. Therefore, modeling spatially separated but temporally correlated behaviour must rely on location-invariant methods. To this end, co-occurrence provides a good way of binding features that are not necessarily spatially proximate or defined by some clearly distinguishable topographical configuration.

The other advantage of co-occurrence across temporally correlated activities is that there does not need to be an explicit assumption about temporal ordering or synchrony. Often in natural scenes, in particular, in surveillance video, there is much human-human interaction. This can vary greatly depending on scene topology as well as the gaze direction of each person. When we are trying to understand whether they are engaging with each other, finding some

loose temporal synchrony is a sensible way to model the commonly observed cause-effect phenomena which is apparent in interactive activity. That is, the order in which the co-occurring features occurred is not necessarily relevant to whether two people are interacting or engaging with each other.

1.4.3 Finding interactions and group behaviour

Finding interactions or salient group behaviour is a challenging but interesting problem. Most human activity recognition work has tended to rely on recognition of activities of individuals. However in natural surveillance scenarios, multi-person activity is usually of more interest since someone is more likely to be in immediate and preventable danger from their external environment. In Chapter 5, human-human interaction is addressed more specifically by representing the dyadic relation between a pose configuration of two potentially interacting people. In addition, the pose configuration or shape information of each person is represented in a form that finds the correlation between different configurations of spatially separated, vertically distributed parts of the body. Using a hierarchical framework to represent multiple layers of temporally correlated features provides a rich framework for classifying human-human interaction.

1.5 Thesis Outline

The following chapters of the rest of this thesis are outlined as follows where the novel contributions are indicated:

Chapter 2 describes the context in which the work presented is set. Themes that are covered will trace the steps between raw data to intelligent inference from it. The main themes will be pre-attentive feature extraction, saliency, spatial and temporal feature binding.

Chapter 3 provides an initial study of how the problem of gesture recognition can be approached. While this is a more specific problem domain, the solutions presented provide motivations for the work presented in the rest of the thesis.

Chapter 4 addresses the problem of finding salient features in video whilst in the presence of non-stationary background. It describes a method of quantifying temporal saliency using an extension of the scale saliency algorithm of Kadir and Brady (2001). The method is shown to be effective at differentiating noisy moving background from salient moving foreground. In addition, the proposed method is able to select both a salient spatial and temporal context in which to consider regions of interest.

Chapter 5 explores the idea of feature binding for understanding interactions and temporally correlated behaviour between objects in a scene. Here, the method described in the previous chapter is used as a building block for finding co-occurrences in a neighbourhood of spatial locations as well as spatially separated but temporally correlated locations. Here, the proposed algorithm is able to detect salient interactive events and highlights the power of using co-occurrence to recognise unusual activities. These are activities which may not have been detected if no explicit correlation between regions of interest had been represented.

Chapter 6 investigates feature binding specifically for interpreting human-human interactions from spatially separated people. The novelty of this approach is that rather than using a supervised approach to represent the kinematics of the human body, a simple pose-based configuration model is used to approximate limb positions of each person. In addition, the idea of explicitly representing the interrelation between two potentially interacting people using canonical correlation analysis is presented. Levenshtein distance is then performed on a multi-hypothesis feature trajectory space in order to classify three different classes of interaction and non-interaction. Further modifications are also made to the spatial saliency algorithm so that it can be applied to the spatio-temporal domain for the detection of salient human forms. The results highlight the limits of Kadir and Brady's original spatial scale saliency method and offer some improvements to it.

Chapter 7 draws conclusions from the findings of the thesis. The value of the work presented in here will be debated and future directions will be discussed.

1.6 Publications

The work presented in this thesis has been published previously, in the papers listed below.

- H. Hung and S. Gong, "Classification of human interaction from a distance using salient body behaviour modeling", Proc. SPIE Conference on Optics and Photonics for Counter-Terrorism and Crime-Fighting, Stockholm, September 2006.
- H. Hung and S. Gong, "Detecting and Quantifying Unusual Interactions by Correlating Salient Motion", Proceedings of the IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS'05), Como, Italy, p46-51.
- H. Hung and S. Gong, "Quantifying Temporal Saliency", Proceedings of the British Machine Vision Conference (BMVC'04), Kingston, UK, p727-736.
- H. Hung and S.Gong, "A Bottom-up Approach to Quantifying Saliency in Video", submitted to Pattern Recognition.
- H. Hung and S.Gong, "From Temporal to Spatio-temporal Saliency", submitted to Pattern Recognition.

2 Background Review

2.1 Introduction

Pre-attentive extraction and representation of salient features is a desirable aspect of automated video scene understanding. Something which is defined as salient can be 'strikingly conspicuous', 'prominent', 'outstanding', and 'observable' [1]. Each of these words describes the idea that something salient can be readily noticed or observed, immediately apparent, stands out, attracts and/or holds our attention and is unusual or extraordinary. The meaning of saliency in the scope of this thesis is very similar in the sense that we are trying to find elements of a scene that are readily noticeable and that may attract and hold our attention. In addition, the salient features should be meaningful in some way. That is, it would be preferable if the extracted feature could be explained on a semantic level.

Given that there are hundreds of thousands of pixels in an image sequence, computing on all these pixels, whilst ensuring computational tractability is impossible. Moreover, there is likely to be considerable redundancy in parts of a scene, such as areas of background. Parts of an image sequence that are considered to be background tend to contain very little information about a scene in proportion to the number of pixels that represent it. Performing much computations on these areas is clearly very inefficient. Therefore, being able to extract one or many regions that hold our attention due to being unusual and immediately observed, is certainly desirable.

From a human perspective, many of the salient features we can see at first will be extracted using pre-attentive mechanisms, which will be described in more detail later. For now, let

[1] The American Heritage Dictionary of the English Language, Fourth Edition, 2000, Houghton Mifflin Company

us assume that these represent features with some local spatial interest such as areas of high contrast or edges. However, once these features have been extracted, they do not have any representative meaning and must be bound together so that a more compact representation can be formed. At this stage, we might discard some of the originally selected features due to the prominence of particular groupings of other features. This selection process via feature binding has been suggested by Gestalt Theory, which shall be explained in more detail later.

Once feature binding has occurred, we are still left with much data to digest that may not be relevant to us. Relevance here refers to visual stimuli of lower priority since our eyes are continually bombarded by such large amounts of data which in practice cannot all be interpreted (Itti and Koch, 2001). Discriminative selection of those visual stimuli which are of interest to us, ultimately reflects much about the internal state of mind. This was demonstrated by Troscianko et al. (2001) who showed 20 second image sequences of surveillance footage of interacting people to subjects with and without experience in surveillance video monitoring. In each case, the video stopped before any potentially criminal behaviour occurred and the subjects were asked to predict whether any antisocial or criminal behaviour might occur. Overall, correct predictions were made in most cases. Furthermore, they also found that professional monitoring staff had a higher false positive rate due to the precautionary nature of their profession. This example highlights the difficulty of trying to find salient features from an image sequence and how task driven attention can be. Task-driven attention has also been exploited by marketing experts to design spatial arrangements of different products on supermarket shelves where they found that customers tend to fixate more on brands rather than prices (Chandon et al., 2007).

If so much of what is considered salient depends on the context, how much of the context can be decided from bottom up approaches rather than prior knowledge? Can the context be selected automatically? From the perspective of computational complexity, we wish to extract the maximum information out of any content present in the data before applying top-down models.

A good way to describe saliency mathematically is in terms of statistical unpredictability. The context in which this is defined can be seen as the point at which something becomes

unpredictable, given a model of predictability. Saliency is intrinsically linked to context since something can only be unpredictable if it is accompanied by a model of what is predictable. Creating this model of predictable behaviour requires us to select the context around which it should be built. The fundamental questions which will be addressed in this chapter are:

- How can we extract useful and meaningful features from a scene?
- How can we represent these features in a suitably discriminative or readily measurable manner?
- Once these features are extracted, how can we make sense of the large quantity of features that are extracted by binding them together?
- Once groups of features are bound together, does the relative saliency of different groups of features change?

If we approach the problem of trying to extract the meaningful information from a video sequence (rather than single images) in stages, we may choose to start at pixel level, extracting motion patterns from the temporal dynamics of it. Once such spatio-temporal regions of interest have been identified, it is desirable to represent these features at a higher level so that we can approach a semantic-level interpretation of the scene. We intend to model the dynamics of spatially, temporally and/or spatio-temporally correlated behaviour between salient regions and find a feature binding strategy that provides meaningful spatio-temporal structures. Behaviour in this context refers to a set of spatio-temporal motion patterns which have been grouped together under some binding process, where the resultant group can be mapped to a higher semantic meaning than the original disparate parts.

In this chapter, common methods and important contributions towards solving the problem of representing and understanding video will be described and discussed. Starting at the first stage, Section 2.2 describes the motivation behind performing bottom-up feature extraction through biologically inspired approaches. After these low-level features have been extracted, their state is still fairly primitive in the sense that they will tend to represent low level spatial concepts such as edges, corners or regions of interest. They are clearly not enough on their

own to signify higher level semantic ideas of behaviour and activity. This notion can be explained better with an example of what low and high level features might represent in a semantic context. In the case of the low level features, one might describe a feature as 'this is an edge', whereas if these features were bound together, we might be able to say 'these sets of edges combine to represent a person'. Extending this to the temporal domain in Section 2.3, the higher level might be equivalent to 'these changes in the edge over time combine to represent the activity of walking from a single person'. Clearly binding the low level features together is non-trivial and is much dependent on context and what the final application might be. Therefore, feature grouping will be discussed with reference again to the human visual system through the idea of the binding problem in Section 2.4. Finally, various methods for detecting interactive behaviour will be presented and discussed.

2.2 Bottom-up salient spatial feature selection

As discussed in the previous chapter, saliency is a very important aspect of image or video understanding. This is made more apparent when we observe the differences between the human visual system and currently favoured feature selection techniques. We can approach the problem by observing the cluttered nature of most natural scenes. Such scenes limit the human visual system to selecting only parts of a scene that we deem to be significant enough to expend further computational resources. Such pre-attentive mechanisms are a natural and logical means of concentrating computational load for faster interpretation of a scene's spatial or temporal dynamics (Itti and Koch, 2001).

2.2.1 Visual saliency - A pre-attentive function

The simplest example of a pre-attentive biological visual mechanism is shown in Figure 2.1(a). In this figure, what immediately stands out or 'pops' out is the horizontal line. According to Gestalt Theory (Benjafield, 1996), the reason that this line is particularly significant for our interpretation of the image is because we try to find ways of simplifying what we see so that it is easier for us to process the visual information around us. In this particular example, the

horizontal line occurs least frequently amongst a set of surrounding vertical lines. We can say that the vertical lines are like the 'background' of the picture whereas the horizontal line is the 'foreground'. It will be revealed later in Section 2.2.3, how the frequency of occurrence of a particular pattern is a fundamental part of saliency detection.

In the meantime, let us analyse further the concept of visual saliency. Saliency was defined at the beginning of this chapter but being able to find a computationally reliable way of measuring the degree to which something is 'prominent' or 'noticeable' is a non-trivial problem. Clearly some visual part of an image or video can only be so if it is considered relative to other parts of the scene. The inherent comparative nature of saliency also leads to another problem since the comparison must be made relative to some predefined context. The context can either be a local spatial neighbourhood, if we are considering images, or a local spatio-temporal neighbourhood for image sequences. However, this is just local context and aside from the fact that considering a local neighbourhood is context dependent, there is also a more global context that should be considered.

If we just consider the simpler problem of an image, then the global context would involve comparing all the imagery data in the image as whole. However, is something less prominent globally if it is less so globally and extremely salient within a local scale? If we add a temporal dimension to our image so that it becomes an image sequence, then the problem is increased further since we must also decide on a temporal context. In this case, defining a global temporal context is much more difficult since in practical terms, we are no longer limited by the spatial resolution of the camera from which the scene has been captured, but the capacity of the storage medium. It is likely that selecting temporal context is the task of calculating whether something is salient within a time interval of seconds, minutes, hours, days, months or even years. Clearly, selecting the context in which to consider saliency is as important, if not more important than the saliency measure itself.

Therefore we come to the difficulty of the problem since saliency can only ever be described as a relative measure. Finding a measure which accurately describes how 'prominent' some regions of the scene are has been approached in many different ways, which will be discussed further in the following sections. For now, let us consider the human visual system and how

pre-attentive mechanisms can be used to select salient phenomena.

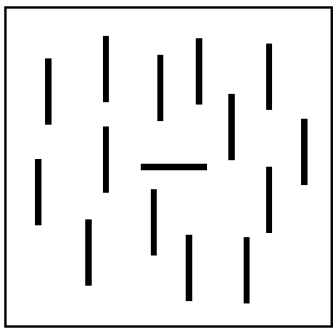
In the human visual system, it is very important to use pre-attentive mechanisms to select regions of interest in order to concentrate computational resources since the optic nerve is only able to transmit signals to the brain at a limited speed (Itti and Koch, 2001). The 'pop-out' phenomenon, described above, is not the only part of the human visual system that tries to simplify visual information in an efficient manner. There are other more sophisticated feature selection schemes that are used for problems of greater complexity such as the visual search task shown in Figure 2.1(b). Neisser's paper on visual search (Neisser, 1964) defined the human visual system as a series of mechanisms arranged in a hierarchical structure. He found that when volunteers were asked to find letters from a list of words, the process favoured searching for distinctive *attributes* of the letter rather than the letter as a whole; finding the letter 'Z' amongst round letters was much easier than finding it amongst angular letters. Such simplification of the search task implies that the pre-attentive stage tends to focus on extracting basic features such as orientation, colour, or motion where a pop-out effect is easiest to identify.

2.2.2 Interest point detection

The idea that the biological visual system picks out basic features such as orientation information and scales was first proposed in the 1960s (Hubel and Wiesel, 1959, 1962). This was then proposed as suitable inspiration for computer vision problems in the 1970s by, Horn (1971) and Binford (1981) who devised a method of finding lines from images of polyhedra under different lighting conditions. Later on, Granlund (1978) also addressed the idea of how to extract suitable features from images using a hierarchical Fourier based technique.

A more common technique, however, extracts information from images using orientation filters (Freeman and Adelson, 1991; Lindeberg, 1998a). Such methods apply various orders of Gaussian filter oriented at different directions in order to locate and quantify the orientation of edges within a scene. This is an effective way of defining edge orientations within an image but does not necessarily define the relative degree of importance of salient regions in a scene.

Methods that are most commonly used to extract features for tasks such as tracking,



(a)

ODUGQR	IVMXEW
QCDUGO	EWVMIX
CQOGRD	EXWMVI
QUGCDR	IXEMWV
URDGQO	VXWEMI
GRUQDO	MXVEWI
DUZGRO	XVWMEI
UCGROD	MWXVIE
DQRCGU	VIMEXW
QDOCGU	EXVWIM
CGUROQ	VWMIEX
OCDURQ	VMWIEX
UOCGQD	XVWMEI
RGQCOU	WXVEMI
GRUDQO	XMEWIV
GODUCQ	MXIVEW
QCURDO	VEWMIX
DUCOQG	EMVXWI
CGRDQU	IVWMEX
UDRCOQ	IEVMWX
GQCORU	WVZMXE
GOQUCD	XEMIWV
GDQUOC	WXIMEV
URDCGO	EMWIVX
GODRQC	IVEMXW

(b)

Figure 2.1: (a) An example of the 'pop-out' effect. The human visual system selects the horizontal line in the middle of the frame using a parallel search mechanism. This figure is reproduced from (Williams, 1999). (b) Two columns of text showing the time differences in finding the letter 'Z' amongst round letters and angular letters. Reproduced from (Neisser, 1964).

object recognition (Lowe, 2004, 2001), or 3D reconstruction (Davison, 2003) tend to be corner detectors (Harris and Stephens, 1988; Pikaz and Disntein, 1994). The objective of these geometric interest points is to provide more robust locations of specific points in an image that are easily located under view or illumination changes.

One of the more popular interest point detectors is the Harris corner detector (Harris and Stephens, 1988) since it is highly invariant to rotation, scale, illumination variation and image noise. This method relies on creating an autocorrelation matrix of the pixel neighbourhood. Interest points are those with low autocorrelation between a central pixel and its surrounding neighbours. In addition, each interest point is given a value corresponding to the 'cornerness' of the region. This technique only works well with images when there is little clutter or noise but a more robust matrix based method was devised by Shi and Tomasi (1994) for tracking point features where interest points are successively removed if local neighbourhoods of pixels do not match enough between frames. This inherently leads to the selection of feature points that are robust to affine transformations and noise but also display significant degrees of motion between frames. The method was developed specifically as a feature extraction technique for the purposes of tracking. However, the method suffers if objects of interest have little or no textured features around which discriminative measures can be formed. Moreover, interest points are those that provide consistent matching over an image sequence, regardless of how salient the point is over time.

Using consistency as a measure of saliency was also adopted by Laptev and Lindeberg (2003) in their work on space-time interest points. This method selected interest points based on whether motion turning points were consistently repeated over time. Models of interest points associated with different classes were collected from scenes of a person walking against an uncluttered background. Test sequences involved using scenes with cluttered and moving background. The method was able to detect and extract a moving person from the cluttered scene and also estimate their pose at each frame based on the silhouette of the closest model in the training data. However, under more cyclic background variation such as rustling tree leaves, it is likely that the method would become sensitive to a large number of local turning points, which would not necessarily be salient. Their method extracted a walking person from

the test sequence where there was a fast moving car in the background. But the method did not try and extract interest points where a cluttered dynamic background exhibiting periodic behaviour such as rustling trees was present. Moreover, the time period around which the saliency of the interest points is defined was manually configured since the method assumes that the space-time interest points will be periodic. Therefore the temporal context around which the features are found is not selected automatically from the context of the video sequence but from a manually selected time interval which would clearly be different from person to person and also under the execution of different speeds of walking.

Moving away from interest points, Lowe developed a type of region-based feature descriptor (Lowe, 1999) using the Scale Invariant Feature Transform (SIFT) which could represent accurately, a much larger number of regions of interest. Interest points based on extrema of a difference-of-Gaussian (doG) function over the image are selected over increasingly blurred versions of the same image. The blurring in effect, works as a method of increasingly adjusting the scale by which the measurement is taken leading to a scale invariant representation. Using this method, a multi-scale strategy is used to identify regions of interest. This makes it less sensitive to local variations in 3D projection or affine changes since features that are salient over many scales are more robust to generalisation. Therefore, the method has advantages for both object tracking and recognition tasks. However, the SIFT is close to a region descriptor rather than detector and thus relies on finding highly textured regions of interest. It is easy to see that it would not be able to quantify the saliency of the regions of interest or select them based on a global context since it is primarily searching for regions of local spatial change.

More recently, real-time methods of finding and tracking interest points for applications such as robot guidance has led to a push for more efficient computation by combining tracking with adaptive interest point detection (Davison, 2003, 2005; Rosten and Drummond, 2006) in a top-down approach. Schmid et al. (1998) provides a comprehensive guide to different types of interest point detector for further reading.

A common disadvantage of all these techniques is that although the methods are able to pick out interest points fairly robustly, the points are only of interest in a local sense. Without a measure of how salient the region is, it is impossible to select regions based on more global

ideas of saliency. Furthermore, they do not address the problem of how to group the features together into more coherent objects. This is particularly the case since interest points are corners or turning points in either space or time, which can only ever represent part of an object. If there are multiple objects in the scene, with an unknown number of feature points associated with each, grouping them together becomes a non-trivial problem in particular, when they are close or overlap spatially. Moreover these points of interest are selected by a binary process, which tends to lead to either over selection or under selection regardless of the global context within which the points are selected.

It would be rather more convenient if salient homogeneous *regions* of interest could be extracted in one step with a reliable measure of how salient or prominent the region is relative to others. Moreover, most of the schemes described above do not consider how the temporal evolution of interest points might render some less salient than others due to their frequency of repetition. From a practical viewpoint, without more efficient feature selection in spatially cluttered image sequences, over-selection could become a problem and may make these methods intractable for interpreting cluttered outdoor scenes where dynamic background changes in lighting exist.

2.2.3 Spatial saliency detection

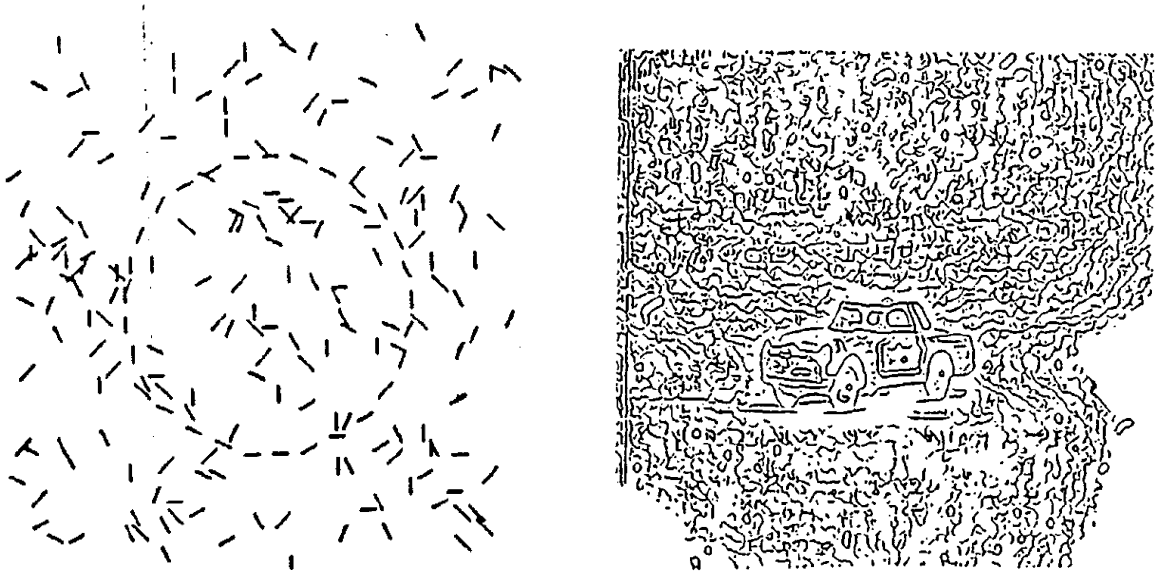
The main difference between interest point detection and spatial saliency detection is the method by which saliency is considered. While interest point detectors try to select local changes in the spatial data, saliency detectors perform a semantically higher level of extraction of what is fundamentally important in a scene by selecting quantifiably discriminative regions of interest. That is, the features can be extracted in such a way that their representation is closer to a symbolic or semantic meaning that can be related to everyday objects. It is closer to the idea of visual 'pop out' that is exhibited by the human visual system and also the Gestalt theory of feature binding, which considers how features can be grouped together to form some higher meaning than the separate parts. That is, saliency detection tries to extract what is not only locally significant but also globally so.

In the early 1980s, Koch and Ullman (1984) proposed the idea of pre-attentive visual

attention mechanisms that could be implemented by hardware. Ullman went on to address the idea of structural saliency in images (Ullman and Shashua, 1988) by defining curved structures to be more salient. Feature extraction methods at that time tended to rely on orientation results which were good at finding edges but inferring regions from this was not a simple task when given complex images such as that shown in Figure 2.2 which illustrates a clear visual 'pop-out' effect. Using ordinary orientation extraction methods, everything in the image would be considered salient since it would be impossible to tell different line fragments from each other. Ullman devised a structural saliency measure based on variations of curvature of groups of broken lines, where the results are shown in Figure 2.2 where the method was able to pick out the shape of a circle from (a) and the outline of a car from (b). The idea that the human visual system favours round contours or blob-like structures was first suggested by Uttal (1975) when he showed subjects arrangements of dots. He found that when the dots were placed close together as a curved or regular pattern, subjects were more likely to recall them.

One of the seminal works on saliency detection in images was done by Itti et al. (1998), who drew specific inspiration from the human visual system to extract salient local orientation information from images using intensity, colour, orientation and other visual cues. Salient regions could be detected within still images by implementing a hierarchical multi-scale structure with a winner takes all strategy. If more features were required, the feature selection process could be iterated to select the next most salient feature and so on until no more features remained.

As we have seen so far, many spatial feature extraction methods rely on the use of orientation filters (Itti et al., 1998; Laptev and Lindeberg, 2003; Schiele and Crowley, 1996; Chomat et al., 2000; Lindeberg, 1998b). Such techniques are not necessarily effective for extracting salient features since regions that produce a higher magnitude response from orientational filters are largely dependent on the choice of the basis functions which can be rather arbitrary. Using orientational filter responses to identify salient parts in an image would mean that even cluttered background could be extracted readily as a significant part of the scene. The problem can be demonstrated by the example shown in Figure 2.3 where (a) is filtered



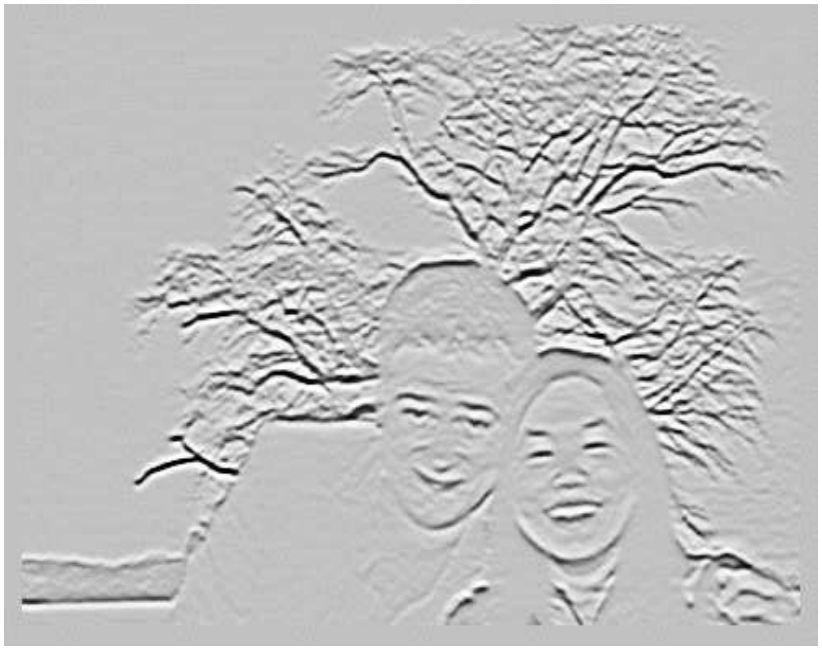
(a) (b)

Figure 2.2: Two images reproduced from Ullman and Shashua (1988). Both demonstrate the effect of visual pop-out but also the idea that this can be emulated using Ullman's structural saliency measure. (a) A circle composed of broken lines surrounded by a background of many broken lines placed randomly at different orientations. (b) An edge image of a car in cluttered background.

with a horizontally oriented Gaussian filter to obtain (b). Most of the high magnitude responses using the Gaussian oriented filter identifies the tree branches in the background of the picture. Since the tree region is very self-similar, it would be sensible to assume that such a region would be considered homogeneously as a single background texture. Therefore what we might consider to be more salient in the image are the figures in the foreground.

One might also argue that implementing a system which shows similarities to the human visual system is not necessary when we are trying to approach automated image and scene understanding where using statistical methods are more preferable. To this end, Kadir and Brady (2001) devised a simpler scheme based on the notion of entropy in information theory. The advantage of using entropy was that it could be equated to saliency by measuring the impurity of a set of data. The higher the entropy of a set of data, the less predictable it is since the information is not very pure and therefore it is more salient. Therefore, as opposed to many interest point detections which rely on arbitrarily chosen basis functions to define

(a)



(b)

Figure 2.3: An example of orientational filter-based feature extraction. (b) shows the effects of just applying a first order Gaussian filter to (a) where the filter was oriented to extract horizontal edges. Although the people in the foreground are clearly of more interest in the picture, it is the cluttered background that contains more higher magnitude responses.

points of interest, Kadir and Brady's scheme provided a much more general approach to image saliency. Using this local generalised context[2] scheme provided an efficient way of both measuring and selecting the region of interest based on the variation of the local neighbourhood. This is very different from the more traditional interest point detection methods which relied on the idea of measuring the local gradient information alone (Lindeberg, 1998a; Lowe, 1999). Therefore, using the methods described in the previous subsections, the noisy image shown in Figure 2.4 would produce many interest points. Surely, a point is only of interest if there are no other similar ones close by.

However, Kadir and Brady found that whilst high entropy could equate to high saliency, problems were encountered with noisy self-similar images. This is shown in Figure 2.4 where three permutations of the same image have identical entropy. The first image shows part of a face, the second has the image pixels ordered by intensity into a ramp, the third has a random ordering of the same pixels and the 4th has the intensity values of the same pixels formed into a circular gradient. In each case, exactly the same pixels are used but their re-orderings lead to contrasting spatial configurations. If the entropy is calculated over increasing scales using a circular kernel centred at the centre of each image, then the entropy-scale characteristics are very different. In particular, the graph on the third row of Figure 2.4 shows an extremely flat entropy-scale characteristic when the pixels in the top image are randomly distributed around the image frame. Conversely, the entropy-scale characteristic of the top image shows a maximum at the scale that corresponds to the size of the iris.

The advantage of selecting the scale of salient regions based on a statistical measure of the data is that local context is an inherent part of the feature extraction process. Therefore the level of spatial homogeneity can in itself, be defined relative to a local context since the entropy measure does not explicitly represent one colour/intensity distribution. This is particularly useful since we would clearly prefer to select regions of interest which may equate to a higher semantic meaning, regardless of how textured or varied its local spatial structure is.

A disadvantage of Kadir and Brady's method is that although salient regions rather than

[2] orientational ordering is discarded in favour of radial ordering

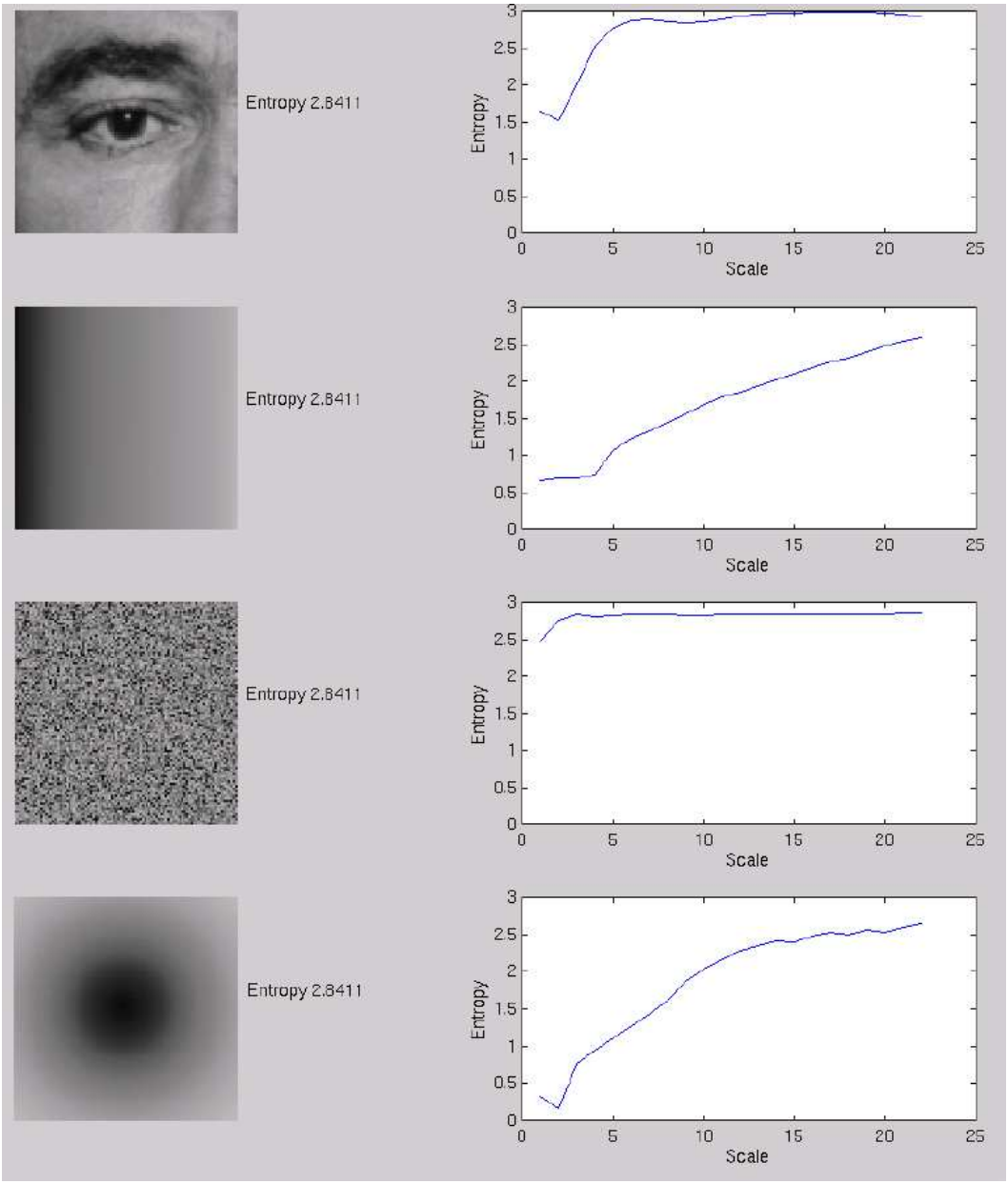


Figure 2.4: Entropy-scale characteristics of the same set of pixels shown at different permutations. Whilst the entropy of the the entire image for each image is identical, the entropy-scale characteristics taken using a circular kernel centred at the centre of each image, show very different variations over scale. This figure is reproduced from (Kadir et al., 2003).

points were detected, there was still no method of selecting globally salient features. While spatially homogeneous regions can indeed be located, the detection of people, for example, is still not possible since people can be composed of many spatially homogeneous regions, some of which are more spatially salient than others. On the other hand, a method of finding salient configurations of spatial data was introduced by Boiman and Irani (2005) who use local shape information to detect unusual variations in configuration and their respective local intensity Probability Distribution Functions (PDFs). They didn't perform any initial feature selection but rather computed the local intensity histogram of a local neighbourhood over different scales and calculated the likelihood of every combination of configurations of local neighbourhoods over an image or image sequence. By exhaustively comparing every possible local configuration, it was possible to detect unusual spatial configurations. However, it is not clear from their experiments whether the methods would work in more cluttered background. They also applied this idea to define spatio-temporal saliency (described later in Section 2.3.2).

2.2.4 Summary

In this section, different methods for selecting and detecting spatially salient features from images has been presented and discussed. While interest point detectors are effective for analysing an image, finding regions of interest brings us close to a higher semantic representation of the data. However, so far, we have concentrated on selecting features in the spatial domain. Clearly, the temporal evolution of spatial features can play a huge role in determining whether a particular region of interest is as salient, given a temporal context. Therefore, temporal saliency and other forms of temporal change will be considered in the next section. The themes will draw from the same basic ideas of spatial saliency detection such as the importance of selecting context but will also consider temporal and spatio-temporal saliency.

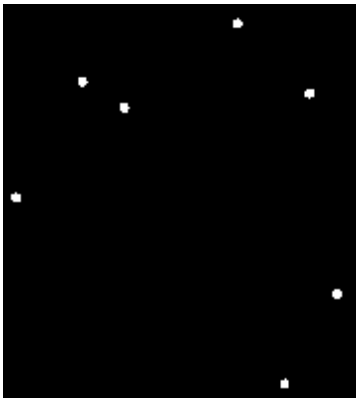


Figure 2.5: An image of something, with the salient parts marked by spots.

2.3 From spatial to temporal saliency

Let us firstly consider how considering temporal evolution can provide a profound impact on the information that can be extracted from a scene. There has been compelling evidence to suggest that selecting a few carefully chosen points from the surface of a deformable object from an image sequence can provide much information about the content of an image sequence (Johansson, 1975; Giese and Poggio, 2003; Wertheimer, 1961).

Johansson (1975) showed that even with a few manually selected points of a person walking, it was still possible for humans to recognise the actions being performed. He presented volunteers with a static image of spots such as those shown in Figure 2.5. However, without temporal change, it is unclear what this image represents. Showing this image as a sequence (see Figure 2.6), the light spots appeared to move. From Figure 2.7, we see that the sequence of spots are actually the positions of distinctive parts of the human body and that the extracted sequences are taken from a sequence of a person waving. There were two different aspects to this discovery. Firstly, it was possible for the human observers to perceive motion of the dots and therefore their association between frames. This perceived motion, which is part of the Gestalt psychology was discovered by Wertheimer (1961) who called it the Phi phenomenon: humans can perceive motion, even if something is not physically moving. He found that when 2 lights, which were placed apart, were switched on alternately, instead of seeing 2 lights being turned on alternately, what was perceived was one light moving from

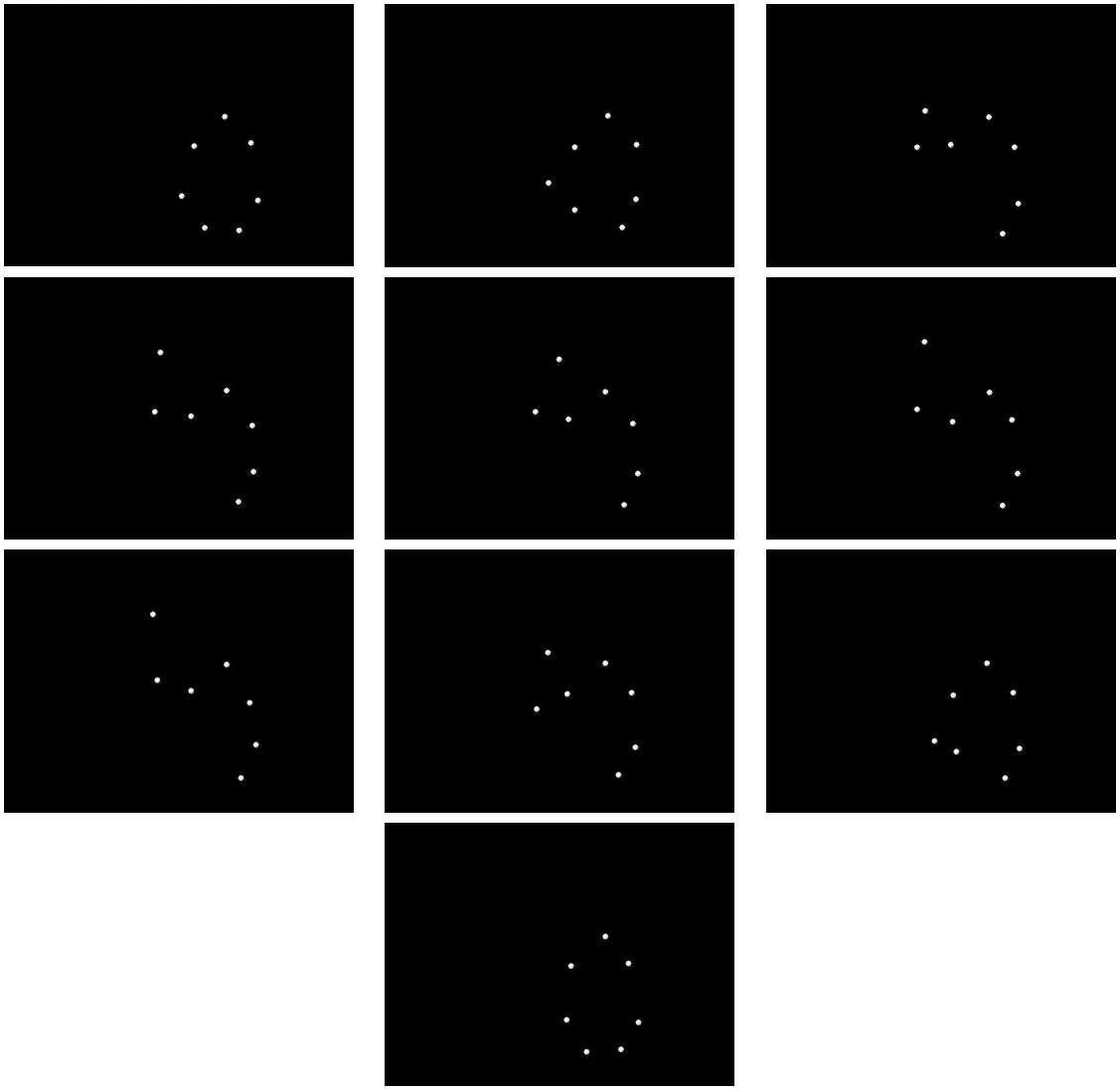


Figure 2.6: Frames of a Moving Light Display (MLD) where lights were placed at a predefined set of locations on a deformable shape.

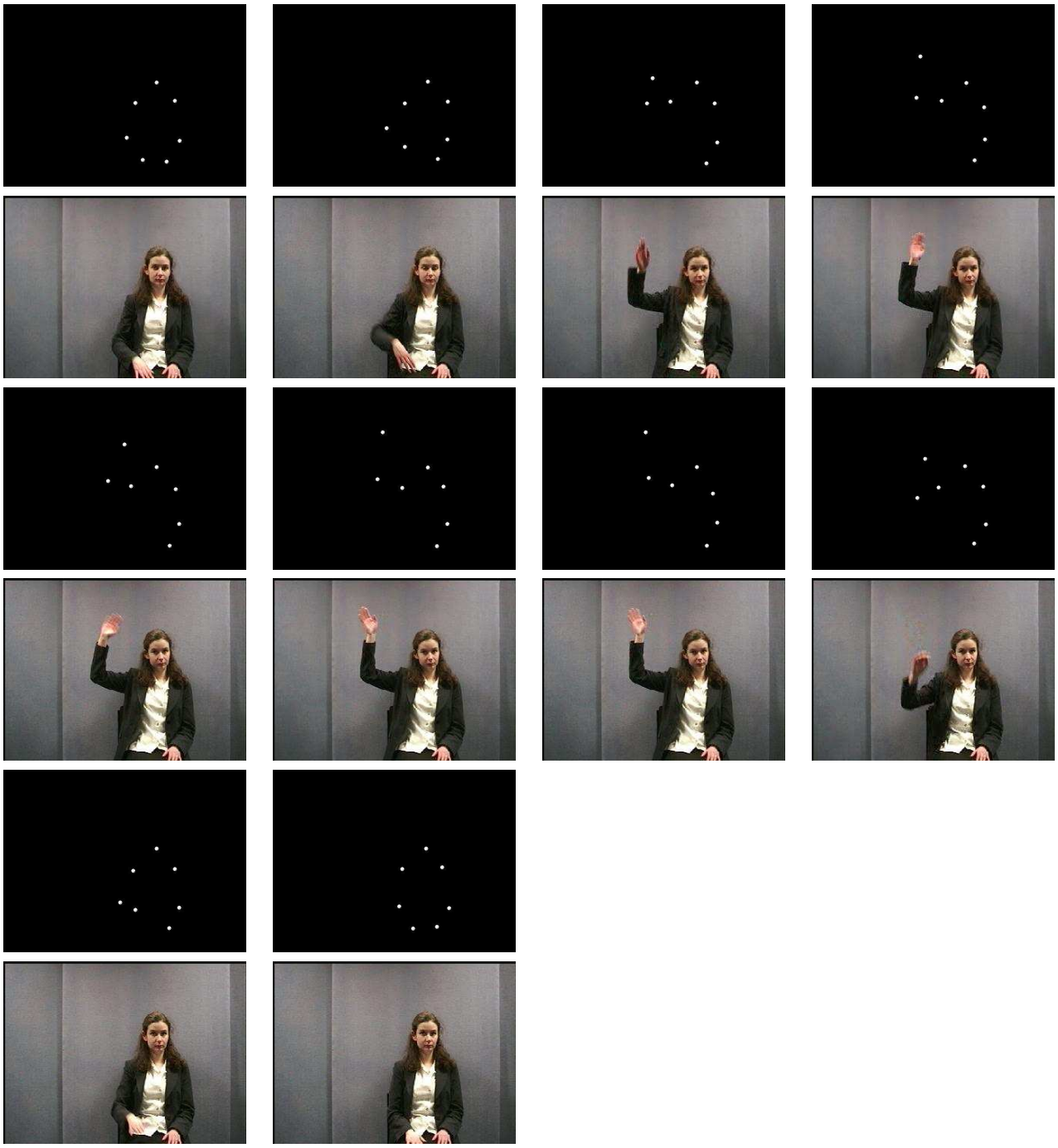


Figure 2.7: Frames of a waving sequence, with their corresponding Moving Light Displays (MLDs). The action is still distinguishable when subjects are shown a sequence of moving images where only white spots corresponding to the joints of the performer are shown.

side to side. Secondly, prior knowledge of the world helped them to see that the sequences were of a person performing a particular action.

Giese and Poggio (2003) carried out some experiments by using moving patterns of dots, where each spot spatially overlapped different parts of the human body during the action of walking. They found that human subjects could identify the walking person easily. What is it about the movement of these spots in Figure 2.7 that makes it clear to us that it is a person waving? It is likely that the motion was anatomically valid and therefore familiar to the subjects. However, if fewer dots were used, is it still clear that the apparent motion of the remaining dots represents someone waving? In this case, the answer lies in which dots remain. For example, if 2 dots that are used to represent a stationary part of the body were removed, it is likely that this will not impact greatly on our ability to perceive the motion. However, if 2 dots are removed from the more significant parts of the motion, such as the hand and elbow of the left hand, then it will become extremely difficult to perceive the motion. This was shown by Giese and Poggio (2003) who found that motion pattern neurons in the brain responded less to point-light stimuli of a person when the sequence was degraded by removing different dots, as shown in Figure 2.8. Neuron activity was reduced to different degrees depending on which dots were removed, thus reiterating the idea that some features are more salient than others and also that maximising the saliency of the chosen features is important.

Clearly an automated system must be able to extract the temporally salient features too. Although all the dots exhibit some motion, finding the dots that cause salient motion is a non-trivial task. We can generalise the dots to salient regions of interest over time. However, in surveillance video, for instance, scenes are cluttered with not only moving foreground but also non-stationary background. Often background motion can be quite erratic and difficult to ignore unless complex models of normal behaviour are applied (Hamid et al., 2005; Zhong et al., 2004; Xiang and Gong, 2005b).

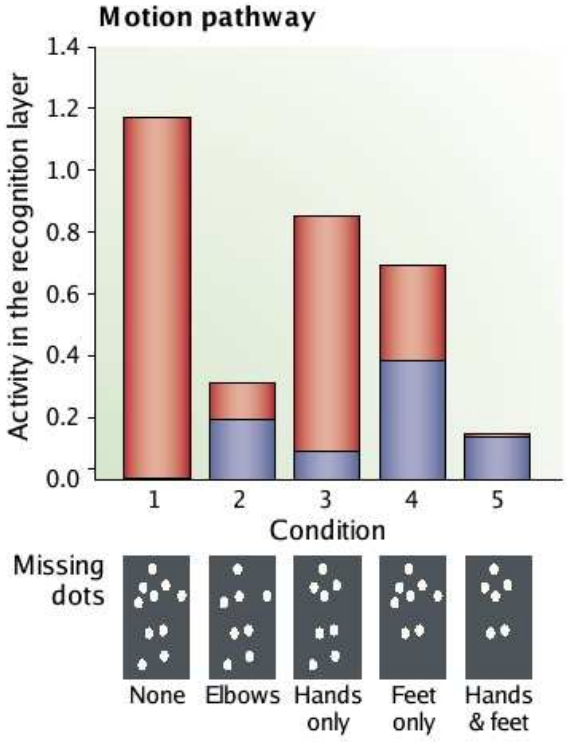


Figure 2.8: Activity of the motion pattern neurons for point-light stimuli that were degraded by removing different dots. Red bars show the activity of the neuron that responds to walking and blue bars show the maximum activity of all motion stimulated neurons. Reproduced from (Giese and Poggio, 2003).

2.3.1 Estimating and detecting motion

The importance of motion is self-evident as a method for extracting meaningful movements from image sequences. A popular technique for measuring motion is through computing optical flow (Bergen et al., 1992; Horn and Schunck, 1981; Lucas and Kanade, 1981; Spinei et al., 1998; Wixson, 2000). These try to find flow fields within an image sequence by matching neighbourhoods of pixels between frames to find directions and magnitude vectors of flow. Estimating optical flow can be organised into 2 stages. In the first stage, local displacement information is calculated and in the second, global constraints are applied to produce a more globally consistent flow field. These 2 stages highlight the importance of integrating some notion of local motion into something more globally meaningful. It is particularly necessary since if motion is only considered in a small region where a contour becomes a single line, the

motion is ambiguous [3]. This is known as the aperture problem in visual motion perception where Salzman et al. (1992) found that different sets of neurons were more sensitive to particular directions of motion and that perceptual judgements about motion were formed from combining the information from all the direction sensitive neurons.

Horn and Schunck (1981) first formalised a method of calculating optical flow where the direction of motion between a neighbourhood of pixels is calculated by minimising the difference in intensity of the patch (and deformed versions of it) over time, given a shift in the horizontal and vertical spatial domain. The task then becomes a problem of minimising the sum of errors in the equation for the rate of change of the pixel intensity over time and applying a motion smoothness constraint. The density of flow vectors is high so that the flow of the inner parts of the moving object could be estimated from the motion of the outer parts of the object. However a higher density of vectors meant that their method was sensitive to noise.

At around the same time, Lucas and Kanade (1981) approached optical flow as an image registration problem. They approached the problem by proposing a more general method of image difference which explicitly ensured robustness by weighting the distance by a normalisation term which made the estimate more reliable if the variation in the local spatial intensity was more linear. Usually, reliable registration was more likely under smaller distances between images where the difference is assumed to be linear. However, this method provided the flexibility to allow differences from larger displacements provided the spatial intensity was suitably linear. The linearity of the spatial intensity could also be approximated by local spatial smoothing leading to even better generalisation. They also generalised the image registration problem further by allowing linear transformations of the images. However, using the pixel intensity to calculate the difference still meant that the method was not very robust to noise.

To this end, Hildreth (1983) used the more locally distinctive feature of contours to estimate more stable motion displacement. This led to better robustness against noise but led to an inherent ambiguity in the local motion direction due to the aperture problem. This was

[3] That is, any motion occurring along the contour cannot be determined

overcome by addressing the overall velocity vector in terms of global smoothness constraints. Gong and Brady (1990) extended this technique through a tighter motion constraint for better estimation of the overall velocity vectors as well as faster computation through parrellelisation of the method. However, with all optical flow methods, the fundamental assumption is that all the motion that is estimated is of interest. Hildreth (1983) tried to improve this by selecting edges from which to estimate the flow vectors rather than just arbitrary patches. However, as we have already seen from the previous section, edges do no necessarily provide the best representation of the most interesting parts of an image or image sequence.

Another way of viewing the problem is that the edges represent unpredictable changes in the local spatial context. Therefore, it seems sensible to use the homogeneity of the image intensity as a means of accumulating a model of solid regions of an image. This method challenges previous assertions about how to find salient homogeneous regions of interest, which was discussed by Toyama et al. (1999) in their seminal work on managing background models. In particular, they address the problem of finding salient homogeneous regions of interest in image sequences when the object motion is relatively small. Hence between frames, only the frontier edges of the object are highlighted as foreground. Figure 2.9 illustrates their approach to finding salient homogeneous moving regions. They start by labeling individual pixels as foreground or background and then decide whether an object is foreground or background. The label depends on whether the difference between the pixel located at position $\mathbf{x}$ in neighbouring frames, I_t and I_{t-1} is greater than some threshold k_{motion}.

$$J_t(\mathbf{x}) = \begin{cases} 1 & \text{if } |I_t(\mathbf{x}) - I_{t-1}(\mathbf{x})| > k_{motion} \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

where 1 represents foreground and 0 represents background. The subset of pixels which overlap each other in consecutive frames and have also been labeled as foreground in the previous frame, as shown in Figure 2.9(c) are then treated as pixels belonging to the moving object. Pixels are grouped using connected components and a group with less than k_{min} pixels is discarded. In such a framework, firstly, we notice that both the region size and pixel labeling are determined by a manually tuned threshold. It is easy to see that the thresholds

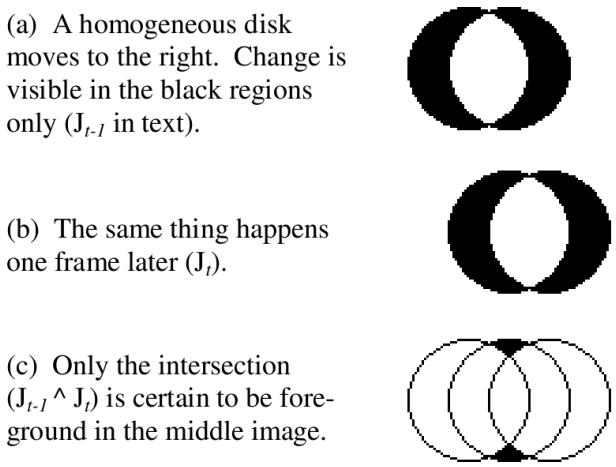


Figure 2.9: Figure reproduced from (Toyama et al., 1999) showing a recommended method of identifying moving homogeneous regions of interest from edge information.

will vary depending on the captured data and hence the method is unstable across a variety of scene data, given the same or similar parameters. Furthermore, the method groups regions on a pixel by pixel basis.

Other methods use receptive fields to calculate motion patterns, which tend to use the biologically inspired idea of multi-scale orientation fields (Adelson and Bergen, 1985; Chomat and Crowley, 1998; Chomat et al., 2000; Freeman and Adelson, 1991) using steerable Gabor filters in spatio-temporal space. The advantage of using steerable filters was that arbitrary angles of orientation of multiple derivatives could be used. Such methods were easier to apply to activity classification methods (Chomat and Crowley, 1998; Chomat et al., 2000; Derpanis and Gryn, 2005) which were highly spatially constrained. However, such methods tend to be computationally costly and provide quite a restrictive motion representation since motion is assumed to be of a linear nature at multi-scales over spatio-temporal space. Furthermore, the method could not select spatio-temporal regions discriminately and would therefore suffer under non-stationary background conditions.

Faster methods to represent patterns of temporal change have been approached through motion history (Bobick and Davis, 2001; Graves and Gong, 2003a; Ng and Gong, 2003; Xiang et al., 2002; Xiang and Gong, 2005a) approaches, based on the accumulation of pixel changes over time. Bobick and Davis (2001) referred to these as Motion Energy Images (MEIs),

which provided a duration invariant method for representing basic human actions. They later devised the Motion History Image (MHI) which added a decay value to pixels so that a representation of chronology was also present. The disadvantage of their method is that it is sensitive to background clutter so experiments were limited to sequences captured indoors.

A method that performs some foreground extraction from the scene was devised by Xiang et al. (2002) who categorised activity in indoor scenes using a Pixel Change History (PCH) measure. Using this method, salient pixels were discovered by applying a threshold to the accumulated change in pixel intensity. These 'pixel events' were then grouped into blobs for tracking and behaviour modeling. The advantage of such schemes is that background/foreground separation for simple indoor scenes with relatively little background motion is also readily achievable within the same framework. This is because accumulating pixel intensity over time provides a natural model of the typical characteristics for that particular pixel. However, blob grouping can become a problem for objects that are spatially connected and requires manual thresholding to remove anomalous foreground pixels.

In general, all the methods described above have tended to rely on the assumption that anything that moves is salient. However, this is clearly not the case in natural scenes. Scintillation from reflections, leaf motion for trees, or just a busy street could be considered as background. Some methods have provided some thresholding to remove some anomalous foreground pixels but relying on manually set parameters leads to a reduction in the generalisation of the method. The important question here is how we can select salient homogeneous regions of motion that can be considered foreground from a global perspective. The other issue involves selecting the salient global scale. This is also of considerable significance but the motion detection methods outlined above do not even try to consider this key element of saliency.

2.3.2 Temporal saliency

In the case of temporal saliency, we are less interested in the small local motion details of an object. Rather, it is preferable to firstly identify homogeneous regions of salient motion so that we can perceive visually what the more important aspects of the motion are. Just

because there is something moving, this does not necessarily mean that it is salient in relation to the rest of the scene. For instance, given the scene is of a crowd of people, is a person's waving arm as significant compared to someone waving in an empty field? From a different perspective, we can also question how salient it is for one person in a crowd to be waving, compared to one out of many in a crowd performing the same gesture. In the first example, we would probably require prior knowledge in order to make a sensible decision about which is more salient. However in the second, it is likely that we can estimate the saliency based on the scene data alone. We can also build a statistical model of how often a person performs a gesture such that if its is observed often, then it is statistically less significant.

Since interest point detection was introduced in the previous subsection 2.2.2, let us start by considering the notion of temporal saliency in a similar manner. As mentioned in the previous section, Laptev and Lindeberg (2003) tried to address this problem by defining space-time interest points as turning points or corners in spatio-temporal space. This is particularly useful for finding unpredictable turning points in the trajectories of moving object. However, their approach was limited to local interest points and did not try to address more global notions of saliency. Rao et al. (2002) also approached the idea of action segmentation by finding turning points in a 2-dimensional projection of 3-dimensional motion trajectories. Both techniques assume that the motion they capture is already salient and meaningful. Furthermore, the spatio-temporal turning points marked regions around which the segmentations of the spatio-temporal volumes should occur rather than volumes themselves.

Therefore, the problem of extracting temporally salient motion still must be addressed. This was achieved to some extent by Oikonomopoulos et al. (2005, 2006a,b) who used an extension of the scale saliency algorithm (Kadir and Brady, 2001) to the temporal domain for recognising human actions. The saliency of the interest point was taken as the mean of the difference in the intensity distributions over two consecutive scales in space and time, which was then weighted by its entropy. Since the measure relies on taking a mean of the change in intensity distributions over two consecutive scales in space and time, there is no disambiguation between spatial and temporal saliency. This is a particularly significant problem since the spatial scale saliency algorithm already has inherent spatial alignment

issues due to the lose of some spatial information in the feature extraction process. They overcome this to some extent by using seed locations which are found using the spatial saliency algorithm, before calculating spatio-temporally salient locations. However, by taking into account the spatial inter-scale saliency, when calculating their saliency measure, the effect of using an initial spatio-temporal seeding method to differentiate regions was reduced. In order to compensate for this, they use an adhoc rule-based approach to cluster spatial neighbourhoods of responses together.

One can also approach the idea of temporal saliency from another perspective by tackling the problem of background modeling. In general, the most commonly used methods of modeling the background tend to either work on a pixel level by modeling the variations in the colour distribution of a particular pixel over time (Stauffer and Grimson, 1999; Wren et al., 1997), or a holistic level by mapping all pixels and their co-occurrence with each other into eigen-space and subtracting the principal components from the representation as the foreground of the sequence (Oliver et al., 2000). Both methods work from opposite extremes where the first ignores any relation between pixels whereas the latter assumes that there is an inherent spatial relation between all pixels in the scene. There have been investigations into using an intermediate technique where blocks of pixels are considered where neighbouring blocks are considered to have less dependence than within block pixels (Seki et al., 2003; Eng et al., 2003). However, for such methods, the size of the block tends to be chosen rather arbitrarily but with quite significant changes to the data. More recently, steps have been taken to try and model more complex background motions such as, sea waves, smoke, fire and moving foliage (Mittal and Paragios, 2004; Doretto et al., 2003). This makes it possible to remove detectable types of background motion but does not address the problem of ignoring previously undefined non-stationary background.

The methods presented show that background modeling is indeed very closely related to temporal saliency. However, in general, background modeling tends to work at pixel levels, looking for regular textures that can be modeled and often incrementally adapted over time. While these are desirable characteristics in temporal saliency, it is even more important to place an emphasis on trying to discover complex foreground patterns of activity

without necessarily being burdened computationally by pixel accurate modeling of cluttered background. Often, the identification of temporally salient objects in a scene require just an approximate indication of homogeneous region of interest, with just a few key features. So, as we have already seen with Johansson's experiments with MLDs, a few carefully chosen spots provides us with all the information we need.

Let us move away from background modeling to a higher level of description of temporal saliency, particularly in terms of unusual behaviour detection i.e. saliency over both space and time. Unusual behaviour detection dictates that something that occurs very infrequently, which is therefore poorly modeled, is likely to be salient or interesting. Such methods tend to accumulate usual patterns of behaviour in order to identify single or relatively few examples of unusual behaviour in feature space that do not match with the rest of the model. Stauffer and Grimson (2000) did this by learning patterns of activity from tracking foreground objects. Foreground objects were picked out by modeling the colour values at each pixel in a particular sequence as a Gaussian mixture model. From this, background subtraction was applied and homogeneous regions were found and tracked. Trajectories were extracted from moving vehicles in a scene in order to accumulate a normal model of motion. Trajectory-based unusual behaviour detection has also been carried out by Johnson and Hogg (1995, 2000) who devised a method of tracking people in car parks. Dee and Hogg (2004) extended this further by also adding goal-directed behaviour based on the topology local to each agent.

More recently, McKenna and Nait-Charif (2004) used priors taken from entrance/exit likelihoods to aid tracking of poorly modeled motion from top-down views of a scene to model unusual behaviour in a nursing home. One might question the importance of camera placement in this method since if the camera had been placed from a side view, then perhaps motion trajectories would not have been so poorly modeled. Camera placement is a significant issue in computer vision tasks. For example, most practical surveillance systems prefer cameras to be placed to maximise area coverage. Often monitoring, by looking at restrictive spaces can lead to cameras being placed in awkward angles for most computer vision systems to cope with. In particular, for cameras placed outdoors, scenes are likely to suffer from representation at very low resolutions in order to increase area coverage.

So far, discussion has focused on a relatively simple representation for discovering unusual behaviour using trajectories of single objects. Clearly modeling trajectories simplifies the problem too much since it assumes that activity is only governed by holistic motion of a single agent. In some cases, tracking is not practical due to occlusion from the surroundings, other agents or scene topology leading to many short trajectories that are difficult to distinguish from one another. Zhong et al. (2004) tried to bypass this problem by matching prototype motion patterns taken from motion histograms with video segments. The method drew inspiration from bipartite co-clustering, a document clustering method use to associate subject matter with emails through word-email co-occurrences. Segments and motion patterns were then clustered in order to find different classes of activity and to identify activities that occurred less frequently. Their approach used quite primitive feature extraction methods that relied on motion cues from relatively uncluttered scenes (i.e. with little occlusion) and strong scene topology dictated much of the typical behaviour.

Hamid et al. (2005) was able to detect unusual activity from modeling complex forms of sequential activity by treating them as 'bags' of events. In this sense, co-occurrences of activities could be formed as N-grams of events. However, their method labeled all activities manually so no automated feature extraction was applied. Labeling activities from the data manually is a more important task than it may at first appear. Clearly, adding expert knowledge or human intervention to a system can be beneficial. However, it can also be detrimental, as shown by Xiang and Gong (2005b), who proved that using manually labeled data made the method less sensitive to discrimination between normal and abnormal behaviour.

More recently, Boiman and Irani (2005) looked at saliency in still and moving images using a shape based model. Their model was able to highlight salient formations based on the frequency of its occurrence relative to a spatial or temporal context. Using a shape based model is a very effective method of modeling local spatial pose variations. This is illustrated in Figure 2.10 where the left image shows typical actions whereas the right image shows an unusual action. The right image also highlights in red, the areas of the frame where salient activity has occurred. While these configurations are indeed salient compared to typical activity, perhaps the more important observation from this example is that the interaction

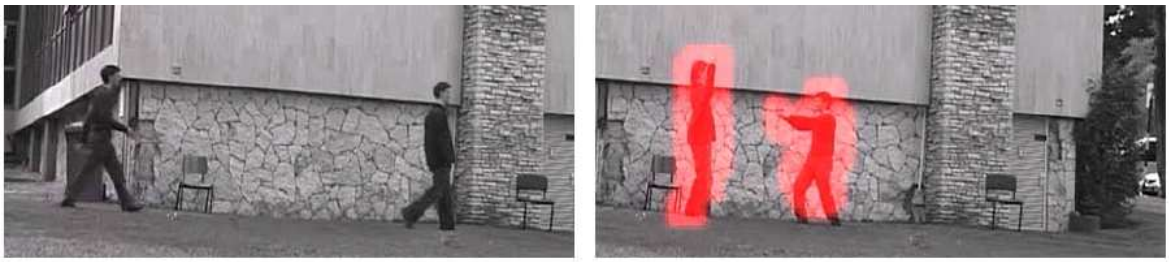


Figure 2.10: Figure containing typical activity on the left and unusual activity on the right. The picture with unusual activity shows regions highlighted in red where salient configurations have been detected. It is important to note that whilst the configurations are indeed salient, perhaps it is the interaction between the figures which is more salient. That is, the fact that both these salient configurations occur simultaneously is more significant than their spatial configurations alone. Reproduced from (Boiman and Irani, 2005).

between the salient subjects is even more significant. For example, if the scene was of people playing basketball, the action of raising arms would not be considered so salient. However, if someone was to walk on using a similar pointing configuration, suddenly any temporally correlated behaviour with this motion would be much more significant. Furthermore, their shape model did not select explicit features but instead extracted all possible combinations of features at all scales at every point in a local spatial or spatio-temporal neighbourhood. It is easy to see that with a cluttered non-stationary background, it would become more difficult to distinguish salient configurations without some prior feature selection.

Of course, one could argue that this is just a case of acquiring the right models for the right context. However, this in itself is a non-trivial task and reducing sensitivity to inaccurate modeling is definitely desirable. More crucially, the more interesting forms of saliency are determined by subtle non-exaggerated changes in behaviour which are not always as apparent as unusual spatial forms such as the interplay of less spatially salient features over time. It is important to note that this interplay can occur at close proximity or over considerable spatial separation and being able to find and quantify temporal correlated behaviour is vital for understanding the temporal dynamics of a scene at a higher semantic level. How objects in a scene interact with each other affects their temporal dynamics and therefore the possible responses from the imagery data. Therefore, one must not disregard how intrinsically linked

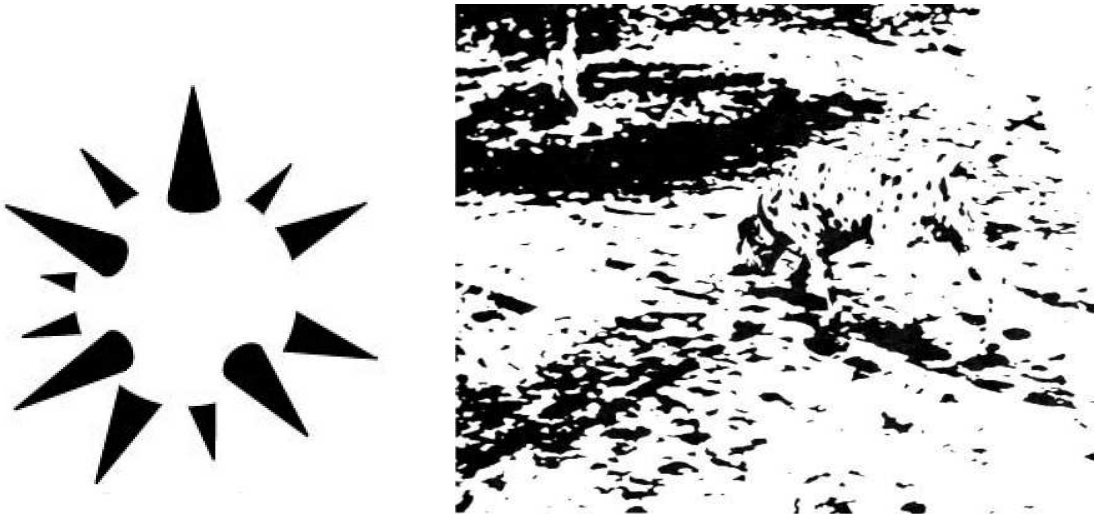
parts of an image can be when it undergoes temporal change.

2.3.3 Summary

The issues covered so far have addressed the problems involved in trying to extract meaningful and salient information in a scene based on spatial and temporal interest. The difficulties that have been highlighted involve selecting the salient spatial and temporal context in which to observe a particularly salient spatio-temporal volume, which is preferably spatio-temporally homogeneous. We have seen that local points of interest in space and time do not necessarily yield globally meaningful features and that finding the right context in which to apply a notion of a larger scale of significance is also a challenging problem. However, there are still more issues of feature representation to be approached, namely, feature binding. This was touched on briefly in the previous subsection where the idea that some behaviour might be considered less salient individually but might be considered more so when considered in combination with other spatio-temporal behaviour. Finding ways to combine features in a suitably meaningful manner is highly context dependent but crucial for image and scene understanding since it leads to a higher semantic representation of the data.

2.4 The binding problem

To understand what is around us, we try to group what we see into easily processable parts. Visual grouping seems at first, a trivial matter if we consider it as a process of clustering or finding some closeness relation in feature space between many neighbouring features. However, as shown in Figure 2.11, where only 2 types of visual binding are shown, there are clearly varied and complex grouping procedures employed by the human visual system in order to interpret and understand the visual world. In Figure 2.11 (a), we perceive a white sphere with black spikes even though all that was drawn were a few black triangles. In this case, all the parts of the image are grouped into the same object through our own prior knowledge about spatial formations. This demonstrates the property of reification. However, in the case of Figure 2.11(b), we must also decide which parts of the scene to focus on and which to



(a) (b)

Figure 2.11: Images demonstrating how the human visual system groups parts of an image together. (a) We perceive a white sphere with black spikes, even though all that was drawn was some black triangular shapes. Image from `http://en.wikipedia.org/wiki/Gestalt_psychology`. (b) Picture of a Dalmatian taken from `http://en.wikipedia.org/wiki/Image:Emergence.jpg` to demonstrate the property of emergence.

ignore. Therefore from all the black regions, a scene can be formed where some of the shapes can be grouped together to form a Dalmatian. This sort of feature binding demonstrates the property of emergence in Gestalt theory.

In cognitive psychology, Treisman and Gelade (1980) proposed a feature-integration theory of attention where they suggested that early visual processes extract features automatically but relations between features is not formed until perception and consciousness are applied. They also claimed that visual attention is necessary for a valid perception of relations between features. This particular idea is extremely relevant to the latter part of this thesis since salient events will be correlated together in order to construct complex, temporally correlated dynamic behaviour between the objects. While bottom-up feature extraction is a significant part of the process towards understanding sequences of complex and highly cluttered scenes, ultimate behaviour understanding must come from inferring meaning from the interaction or from the cause-effect phenomena between objects within a scene.

The idea of feature grouping in Ullman's work on structural saliency (Ullman and Shashua,

1988) was mentioned briefly in Section 2.2.3 where the picture of many broken lines oriented in different directions were used to find a salient structure. The idea of grouping into curved structures is a significant proposition if we consider formations of groups of interacting people. In fact, Kendon (1990) found that formations of interacting humans tend to be based on the geometry of circular formations. Furthermore, for highly engaged groups of people, there is always a convergence to this spatial distribution despite physical perturbations of the formation by one or more of the group or individuals joining or leaving it.

2.4.1 Co-occurrence methods for information binding

Co-occurrence methods have been used a lot in document clustering (Dhillon, 2001) and also texture classification (Elfadel and Picard, 1994; Davis et al., 1979, 1981; Gotleib and Kreyszig, 1990; Haralick, 1979). The advantage of such methods is that they relate features with each other in order to build up more complex structures of grouping and relations. Establishing co-occurrence of features has been adopted for identifying correlations between spatio-temporally correlated features (Stauffer and Grimson, 2000; Zhong et al., 2004). Accumulating co-occurrences of features over a sequence generates a well-defined model of likely correspondences.

In the realms of video saliency, using co-occurrences to quantify the saliency of interactions creates a framework for understanding multi-object behaviours within the scene. This is particularly useful in natural scenes where tracking people and their motions alone is quite a challenging task due to occlusion, capturing the people at low resolution, and stationary as well as non-stationary background clutter. The idea of temporal binding for abnormal behaviour detection was approached by Ng and Gong (2002) who used an accumulated motion history image to find asynchrony patterns in each cell of a frame. The difficulty of modeling patterns of asynchrony was that the resultant vectors were of different lengths, making correlating 2 asynchrony patterns difficult. They found that by using dynamic time warping, it was possible to overcome this problem and model the temporal synchrony of events. However, although temporal feature binding was used, spatial binding was not fully considered and this was demonstrated in their experiments where uncluttered indoor scenes were used.

Other video saliency approaches have viewed the problem in terms of temporal segmentation such that sequences of video are grouped together according to turning points in a time varying feature trajectory (Graves and Gong, 2003b; Gong, 2004). Since both methods approached the segmentation of surveillance video, scenes were cluttered and therefore object selection and tracking was impractical. Instead a pixel-based motion descriptors were used to describe the level of activity in each frame (Graves and Gong, 2003b) or over spatially connected blobs of consistent change in intensity (Gong, 2004). It is still possible to track objects in quite complex scenes but even fitting well-defined object models are not sufficiently robust (Khan et al., 2003; Oh et al., 2004; Zhao and Nevatia, 2003) for interpretation of complex scene behaviour.

2.4.2 Spatial binding through configuration-based modeling

Configuration-based modeling is an important tool for facilitating an automated form of visual binding. Given a group of blobs, we want to know how they can be grouped into something globally meaningful. Shape or configuration based methods have been more commonly applied to object categorisation or recognition tasks (Burl et al., 1995; Fei-Fei et al., 2006; Bar-Hillel et al., 2005; Weber et al., 2000). The huge body of work relying on shape based methods indicates the success of using relative location cues for understanding and interpreting spatial data. We have also seen previously in Figure 2.10 that Boiman and Irani (2005) used spatio-temporal cues to detect spatio-temporally salient actions. Boiman and Irani (2007) have also used similarity between configurations to select configurations as well as scales of commonly appearing spatial patterns in the data. However, finding suitably discriminative configuration patterns means that the scenes need to be less cluttered.

The work of Johansson (1975) has led to many methods that model human action in terms of an abstraction of points or shapes based on the human body (Sherrah and Gong, 1999; Park and Aggarwal, 2003; Ali and Aggarwal, 2001; Efros et al., 2003). In order to bind the moving points together, graphical models have been used to represent the human body. These range from more complex 3-dimensional kinematic models of the human body for body part tracking (Ong and Gong, 2002) to simplified 2-dimensional skeletal forms (Ali

and Aggarwal, 2001) . However, for most classification tasks this is unlikely to be a practical solution since high resolution images are required with little or no occlusion to the body parts. At a slightly simpler level, skeletal graph models of the human body have been used with robustness to pose changes or occlusion of some limbs of the body (Efros et al., 2003; S. Gong, 2002; Sherrah and Gong, 2000; Song et al., 2003). And at the other extreme, highly simplified graph models have been used for action segmentation (Ali and Aggarwal, 2001) or for identifying spatially plausible regions for different parts of the body (Park and Aggarwal, 2003). The disadvantages of configuration-based modeling is that it only considers the idea of binding information spatially. For video analysis of salient behaviour, temporal binding must also be considered.

2.4.3 Interaction detection and recognition

Usher and Donnelly (1998) used psycho-physical experiments to show that, asynchrony can be used to bind events within a time interval to ease computational load compared to binding highly spatially separated salient phenomenon. Therefore, it follows that temporal binding is a significant part of automated scene interpretation and more specifically, interaction detection. Being able to interpret the interplay between objects within a scene leads to more sophisticated contextually driven understanding of the spatio-temporal dynamics. The context-driven aspect of interaction detection is important since it minimises reliance on over-specific expert knowledge that is usually required by top-down models. For instance, much antisocial or criminal behaviour, which is likely to be captured on CCTV is caused by human-human interactions.

One of the areas of information binding that is difficult to capture, yet highly salient, is the notion of interaction between moving objects. For example, in the case of monitoring people and vehicle activities captured by CCTV systems, the physical constraints of surveillance cameras limit the view angle and position that they can be placed at. This can often lead to important parts of a scene being occluded, beyond the boundaries of the field of view, and/or represented by rather low resolution. Under such circumstances, robust object tracking becomes extremely difficult. Xiang and Gong (2006) overcame this problem by trying

to model the temporal correlations of patterns of activity by learning the structure of a dynamic graph. However, techniques that rely on dynamic graphs are limited to a well defined grammar around which states can be discriminately identified. Therefore, a well populated data set is required to discover the temporally correlated structure from the data.

Understanding scene activity in terms of interactions allows us to become desensitised to these limitations by looking for correlated changes in time. That is to say, most interactive behaviour is a cause and effect process so many forms of interaction are likely to be caused by reactions of the objects to its surroundings. Such reactions tend to be behaviour that differs from the local spatial, spatio-temporal and/or temporal usual patterns of motion and can therefore be a good indication of when an object is likely to be reacting to external stimuli from its surroundings. To qualify this salient change or event, one must search for other events within a local temporal neighbourhood.

Many of the more common interaction detection methods are applied to video conferencing or multi-modal applications (Hakeem and Shah, 2004; Gong, 2000; J. Sherrah, 2000; Zhang et al., 2004) but such environments are highly constrained and lend themselves well to very context specific examples. In particular, Hakeem and Shah (2004) recognised the need to look for interactions within a local temporal neighbourhood by using events triggered by the recognition of trained actions in order to determine whether to look for other events within a local temporal window. This relied on the events to be triggered by prior models, which is practical for the constrained environments of video conferencing but less so in outdoor surveillance scenarios.

Interaction recognition tasks has been carried out in less constrained environments with action-based interaction recognition (Hogg et al., 1998; Park and Aggarwal, 2003) but these tended to rely on highly detailed sequences of people, taken from a side view and at close range so that gestures and gait are mostly unoccluded. Many methods create a graph model of the human body (Park and Aggarwal, 2003; Gong et al., 2002) or a shape model of the silhouette of the person (Hogg et al., 1998) where test and training sequences are taken under artificial indoor scenes.

Work on more realistic scenarios has involved looking at interaction of people or objects

based on the displacement of their centroids. Oliver et al. (2000) addressed this problem by modeling the trajectories of synthesised behaviour of humans in a courtyard using coupled Hidden Markov Models (HMMs). Galata et al. (2002) did something similar for modeling car to car interactions on a busy motorway by learning variable length HMMs (VLHMMs). They applied VLHMMs previously to human-human interaction (Galata et al., 2001) which could then define a relevant temporal context in which to consider a particular interaction sequence. However, their feature representation used an active shape model to extract the silhouette of the interacting people so motion overlapping the body was not detectable. With all these techniques, there was the implicit assumption that interactions must occur as spatially proximate locations. This is perhaps not practical when considering natural outdoor scenes where people can interact with each other as long as there is line of sight.

The importance of multiply connected cause and effect behaviour was addressed by Gong and Xiang (2003) who used Dynamically Multi-Linked Hidden Markov Models (DML-HMM) to connect relevant hidden state variables across multiple temporal processes. More explicit treatment of such temporally correlated behaviour of human to human interactions which occur frequently in surveillance data, however remains an unsolved problem.

2.4.4 Summary

In this section, the need and challenges of feature binding was discussed. Carefully selected feature binding has been shown to have a beneficial effect upon the interpretation of a scene, even with very simple image feature extraction methods. More specifically, many of the examples of practical video footage from sources such as surveillance sequences tend to involve much human-human interaction, with potentially antisocial or criminal outcomes. In such cases, finding a suitable way to model the temporally correlated behaviour of individuals becomes an effective method of separating those that are interacting from those that are not. This is particularly useful in scenes with multiple people where classification of activity would be much simpler if it were possible to identify who was interacting with whom first. In contrast, if interactions were only identified by searching for particular actions or gestures from individuals before searching for potential interactions, then the task becomes more

computationally complex, involving more exhaustive computation of all potentially plausible groupings.

2.5 Summary

In this chapter, a number of challenges have been identified within the realms of video behaviour understanding. At the first level, we must try to extract as much meaningful information from a scene by identifying salient regions of interest. Most methods of feature selection rely on interest point detectors, which are useful for tasks such as three-dimensional reconstruction or object categorisation but become less discriminative in extremely cluttered images where extracting just the globally salient features would be more computationally efficient. Under such circumstances, region-based saliency detectors can provide discriminative and quantifiable features. However, for all these cases the idea of temporal saliency has not been considered. The closest example was the space-time interest points of Laptev and Lindeberg (2003) but points of interest were defined as space-time regions which exhibited repeatable behaviour over time.

However, from a statistical perspective, something can only be salient if it is considered unpredictable within some context; local, global or both. Surely if something is periodic, then at some temporal scale, it becomes completely predictable and therefore exploiting the periodic nature of a sequence in order to find interest points defeats the objective of salient feature extraction. Another challenge is deciding what level of context should be applied to the extracted feature. Kadir et al. (2003) and Lindeberg (1998a) suggested using variations in scale-space to identify turning points in spatial feature space. Scale has been defined for video segmentation (Graves and Gong, 2003a; Gong, 2004) but ignores potentially salient regions within the spatial domain. Background modeling can also be viewed as a form of spatio-temporal saliency extraction. However, for this case, foreground features tend not to be selected based on their spatial homogeneity and require pixel grouping, which can be challenging in particularly cluttered scenes. Therefore, the extraction of spatio-temporally salient homogeneous regions of interest has not yet been approached.

At the second stage of feature representation, the features must be grouped or bound together. At the simplest level, features can be grouped spatially according to their proximity in space. However, temporal binding and spatio-temporal binding add additional complexity to the problem. While many techniques reduce the difficultly of the task by assuming spatial proximity to be as important as temporal proximity, this is unlikely to be the case for many natural scenes where people are likely to be influenced by anything which they are within line of sight of. Xiang and Gong (2006) addressed the idea of spatially separated but temporally correlated behaviour but used very primitive feature extraction methods, requiring a manually set threshold on the level of variation in intensity of a pixel to define how salient it is. The fundamental problem that still needs to be solved involves linking more context specific salient features that have been extracted from the raw video sequences into meaningful complex interactive models of the temporal dynamics of activity.

In the last stage, classification of these temporal correlations is required. Specifically, the problem of temporal binding for the detection and classification of human-human interactions. This has been approached to a certain extent, through indoor classification scenarios but still needs more investigations in order to discover better methods of classifying more realistic interaction scenarios.

The next chapter will try to motivate the importance of salient feature extraction further through some example classification experiments of human gesture using more common feature extraction methods. After this, the following three chapters will address salient temporal feature extraction, temporal binding, and finally human-human interaction classification.

3 Quantifying temporal saliency

Automatic scene interpretation of real-world video footage suffers from the presence of cluttered moving background, occlusion, temporally overlapping motion from multiple or single entities and appearance or disappearance of objects. The advantage of extracting features using a context-based entropy measure is that it can represent a statistical model of the variation of a particular image region over space and time [1]. This model has the potential to separate foreground from non-stationary background regions in a scene and to identify temporally salient patterns of change. Non-stationary background is a particular problem since motion can be approximately periodic but determining the period of a region of a moving image depends very much on context or the *scale* at which the region is observed. Therefore to a certain extent, all types of temporal behaviour exhibit repetitivity as long as we find a long enough time period in which to observe them. However, selecting a reasonable scale to observe a spatio-temporal region is a significant problem and the accuracy of the scale selection directly affects the efficiency of the feature extraction and representation process.

As discussed in the previous chapter, effective salient feature extraction is an important part of scene interpretation. Since it is the first stage of automated scene interpretation, it is important to filter out insignificant information from a scene in order to both reduce ambiguity and concentrate computational resources. Natural scenes in particular contain significantly large proportions of background noise as well as salient information. Knowing how to correctly extract the meaningful information from images and video sequences is extremely context dependent. Whilst context can be drawn from top-down models of a scene, it is preferable, at least in the initial stages of scene understanding, to use the scene

[1] A more detailed investigation of the motivation behind the work in this chapter is provided in the Appendixes

data alone to interpret the significance of particular features so the model can be more generic and scene-independent.

In this chapter, the idea of bottom-up feature extraction from video will be explored. In particular, we examine the concept of temporal saliency and develop a model for quantifying it. The scale saliency algorithm of Kadir and Brady (2001) is extended to quantify temporal saliency in order to extract meaningful temporal events from challenging outdoor scenes. The advantage of their scale saliency algorithm over filter-based approaches is that it is able to assess the saliency of an image from a local neighbourhood of pixels using a multi-scale comparison of entropy values. Such a statistical measure of the impurity provides a contextually rich framework for feature extraction. However, while their method was demonstrated to reliably extract salient spatial features from an image, it cannot extract and quantify temporally salient regions. This is particularly important for video behaviour understanding in cluttered scenes where temporal context helps to discriminate salient from non-salient features.

In the rest of this chapter, I will firstly introduce and discuss the spatial saliency algorithm and then formulate an extension to salient temporal feature extraction in Sections 3.1-3.3. Experimental results using our temporal saliency model are shown in Section 3.4.

3.1 Salient spatial feature extraction

A statistical approach to image description might be more sensible than traditional filter-based approaches since using a probabilistic representation means that no information is thrown away when taking local measurements only. Kadir and Brady (2001) used entropy to describe parts of an image in terms of varying scales in space. Entropy is a good way of representing the impurity or unpredictability of a set of data since it is dependent on the context in which the measurement is taken. Traditionally, discrete entropy is defined in terms of the expectation of a random variable X:

$$\mathcal{H}_D(X) = -E_X[log(P(X))] = -\sum_{x_i \in D_X} log(P(X = x_i))P(X = x_i) \tag{3.1}$$

where the logarithm is usually typically defined to be to the base of 2, and the random variable X exists within the set of values in D_X. Kadir and Brady suggest that high saliency is local unpredictability or high entropy and they redefine the discrete entropy term in Equation 3.1 so that each x_i is an interval of pixel intensities (or bins). Here, the random variable is a probability density function (PDF) of the pixel intensity within a local spatial neighbourhood, approximated as a histogram. More precisely, the entropy $\mathcal{H}_D(s_s, \mathbf{x})$ at a particular spatial scale s_s centred at $\mathbf{x} = [x\ y]^\mathsf{T}$ with horizontal and vertical coordinates x and y respectively is defined by the intensity distribution in the form of a histogram of a local neighbourhood of pixels

$$\mathcal{H}_D(s_s, \mathbf{x}) = -\sum_{d \in D} b(d, s_s, \mathbf{x}) \log_2 b(d, s_s, \mathbf{x}) \tag{3.2}$$

where the histogram is accumulated over a circular kernel of spatial radius or scale s_s, centred at the spatial location $\mathbf{x}$ where $d \in D$ are the intensity bins (b) of the histogram. Entropy provides a measure of how impure, and therefore how unpredictable the intensity distribution of the local area is. The novelty of Kadir and Brady's approach is that features are selected based on the variation of the entropy ($\mathcal{H}_D$) over different scales. By looking at the entropy-scale characteristic of a particular local neighbourhood, we can obtain a representation of the local image structure or local *context*.

The next question one must answer is why context is important. Surely entropy alone provides us with enough information about how predictable the scene data is without the need for further enhancements. However, it is easy to see that entropy is extremely sensitive to variations in the scale of the sampling kernel, or in other words the size of its local neighbourhood. The significance of looking at the variation of entropy over scales to find salient homogeneous regions can be shown with a simple example.

In Figure 3.1(a), 3 different regions of interest have been identified for analysis. Starting from the highest to the lowest, they represent a typical sky, tree and eye region from the scene. These have been enlarged and are shown in Figure 3.1 (c-e). Their corresponding gray-level intensity distributions are shown in (f-h). Within the context of this image, we would expect the intensity distribution of the eye region to be the least predictable or the

most salient. The other two regions are much more self-similar and therefore exhibit more predictable behaviour within a local spatial context. However, if we measure the entropy of all these areas, we find that the tree region has the highest entropy whilst the sky region has the lowest. Statistically, the narrow distribution of image intensities in the sky region implies that it is the most predictable of the 3. However, for the tree region, although the texture is fairly consistent, it has the flattest intensity distribution, implying that it is the least predictable region. We clearly need another type of spatial measure to compare entropy values within different spatial contexts. In this way, we can use the predictability of the entropy over varying scales as well as the intensity distribution of the local region to determine how salient it is. To address this problem, Kadir and Brady proposed that if the entropy were calculated at increasing spatial scales, it would be possible to find salient behaviour in the entropy-scale characteristic. In particular, a peak in the entropy-scale characteristic indicates the salient scale of a region since it indicates an unpredicted change in the image intensity data. The idea of searching the scale at which some measure of the image data peaks was first proposed by Lindeberg (1998a) who applied the idea to filtered responses from a scene. Inspired by this idea, Kadir and Brady further suggested that a saliency measure can be obtained from the scalar product of the entropy and some measure of the saliency of the peak in entropy. Therefore, they define an inter-scale saliency measure $\mathcal{W}_D$ in terms of the sum of absolute difference between the intensity distributions at two neighbouring spatial scales as:

$$\mathcal{W}_D(s_s, \mathbf{x}) = \frac{s_s^2}{2s_s - \Delta s_s} \sum_{d \in D} |b(d, s_s, \mathbf{x}) - b(d, s_s - \Delta s_s, \mathbf{x})| \tag{3.3}$$

where the fractional term is a normalisation factor, which ensures that the area of new pixels that is added at each increasing scale doesn't bias the measure towards smaller values of s_s and b, d, and $\mathbf{x}$ are defined as before. That is, the proportionate increase in pixels used to calculate the PDF of the local spatial area would be greater at smaller spatial scales and hence the change in PDF would also be bigger.

The normalisation term is found by finding the proportionate increase in area between

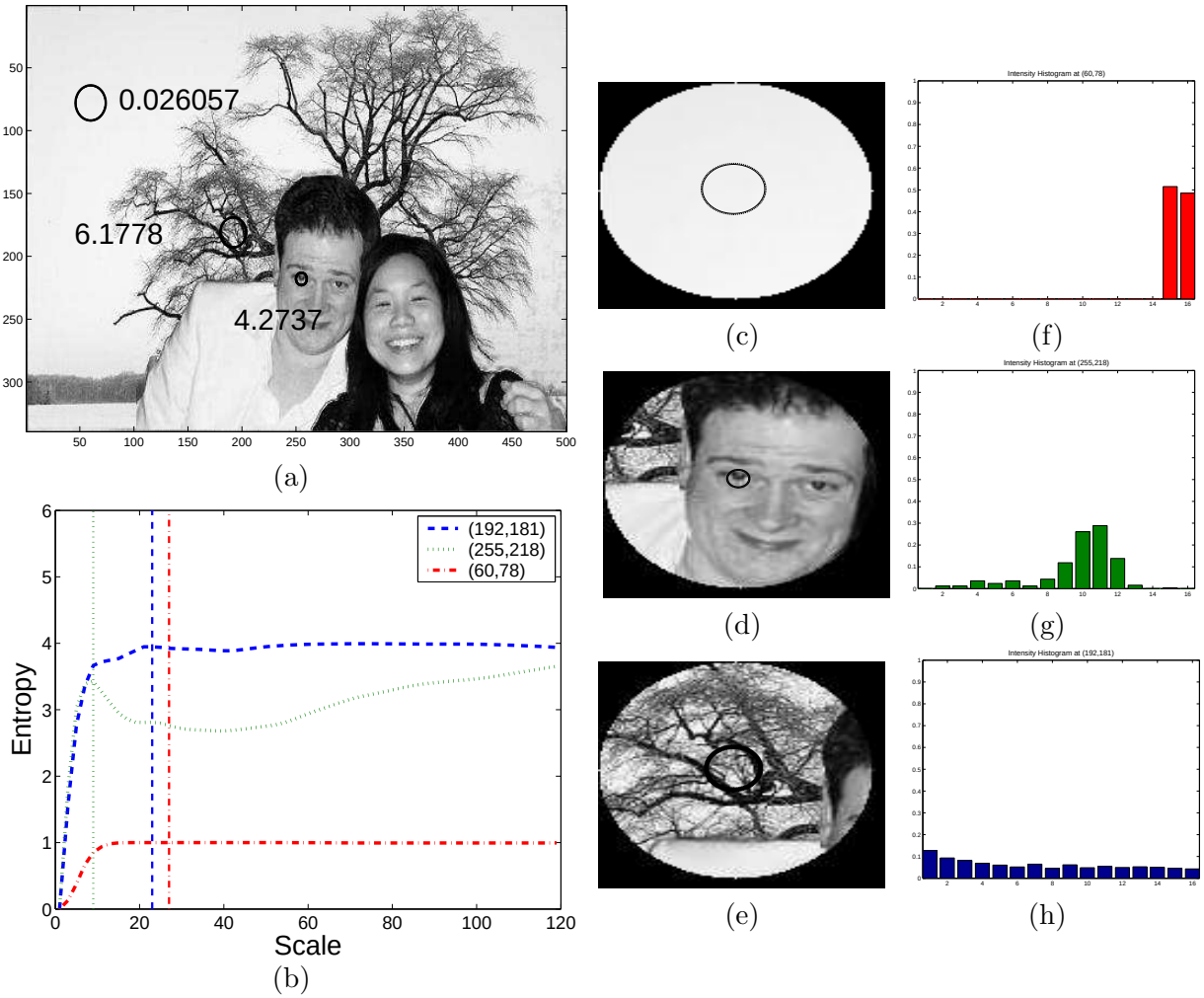


Figure 3.1: An example of the sensitivity of entropy measurements to scales changes. (a) A cluttered image with, saliency values calculated at 3 contrasting points using the spatial saliency algorithm.(b) The entropy-scale characteristics of the 3 regions, where the vertical lines indicate the lowest scale that corresponds to a peak in entropy. (c-e) Enlarged versions of the locations of interest where the size of the circles indicate the spatial scale at which their corresponding entropy peaks. (f-h) The intensity PDFs (approximated by histograms) taken at the scale at which the entropy peaks for the sky, eye and tree region respectively.

neighbouring temporal scales, as defined in Kadir (2002):

$$\frac{\pi s^2}{\pi s^2 - \pi(s-\Delta s)^2} = \frac{\pi s^2}{\pi s^2 - \pi(s-\Delta s)^2}\pi s^2 \tag{3.4}$$

The spatial scale at which the entropy peaks, $\hat{s_s}$, is defined as

$$\hat{s_s} = \{s_s : \mathcal{H}_D(s_s - \Delta s_s, \mathbf{x}) < \mathcal{H}_D(s_s, \mathbf{x}) \wedge \mathcal{H}_D(s_s, \mathbf{x}) > \mathcal{H}_D(s_s + \Delta s_s, \mathbf{x})\} \tag{3.5}$$

where at each location $\mathbf{x}$, more than one peak in entropy over varying spatial scales might exist. Therefore the set of spatial scales with the above condition is defined as $\hat{\mathbf{s}}_\mathbf{s}$. In order to ensure that the peak in entropy occurs at a valid scale where the local intensity distribution is undergoing a genuinely salient turning point, the inter-scale saliency function $\mathcal{W}_D$, is smoothed over scales using a moving average. The saliency, $\mathcal{Y}_D(\hat{s_s}, \mathbf{x})$ at all the scales at which the entropy $\mathcal{H}_D(\hat{s_s}, \mathbf{x})$ peaked is defined as

$$\mathcal{Y}_D(\hat{s_s}, \mathbf{x}) = \mathcal{H}_D(\hat{s_s}, \mathbf{x})\ \mathcal{W}_D(\hat{s_s}, \mathbf{x}) \tag{3.6}$$

where the entropy and inter-scale saliency were defined in Equations 3.2 and 3.3 respectively.

Using this idea of looking at entropy over varying scales, let us refer back to the example shown in Figure 3.1 and observe (b), which shows the entropy-scale characteristic of the three regions. The dashed, dotted and dot-dashed lines represent the tree, sky and eye regions respectively. Although the tree region has a much higher entropy value on average, its entropy-scale curve is much more similar to the sky region. Therefore we can see that the variation of the entropy over scales provides an effective representation of self-similarity. The vertical lines on the graph indicate a peak in entropy for each region, which corresponds to the scales indicated by circles in pictures (c-e). It is evident that the eye region has a more clearly defined peak in entropy than the other two regions. It should follow that it is the most salient of the three regions.

However, Kadir and Brady's model of saliency using an entropy measure between two adjacent scales does not always depict saliency accurately. The saliency values given next

to each of the circled regions in Figure 3.1 (a) show that the most salient region is the tree region, followed by the eye and sky region. As a brief digression, it is important to note here that Kadir and Brady's definition of saliency implies that areas within self-similar spatial or spatio-temporal local regions are considered non-salient. Therefore, the measure should provide maximum discrimination between highly textured areas within a region of background clutter and actual foreground objects.

3.2 A more accurate representation of spatial scale-saliency

In the previous sub-section, it was shown that only taking into account one side of a peak in entropy-scale characteristic does not represent accurately, the saliency of a spatial region. This can be rectified by extending the method to take into account the inter-scale measure at more than two adjacent scales. To this end, we modify the scale saliency by multiplying the inter-scale saliency on both side of a peak. The new term $\mathcal{W}_{D_{peak}}$ measures the inter-scale entropy between the peak scale and the next scale up and replaces the original inter-scale saliency term $\mathcal{W}_D$ from Equation 3.6.

$$\mathcal{Y}_D(\hat{s_s}, \mathbf{x}) = \mathcal{H}_D(\hat{s_s}, \mathbf{x})\ \mathcal{W}_{D_{peak}} \tag{3.7}$$

where the set of scales at which the entropy peaks is still defined as $\hat{s_s}$, and $\mathcal{H}_D(\hat{s_s}, \mathbf{x})$ represents the corresponding maximum entropy values at the spatial position $\mathbf{x}$. The modified inter-scale term, $\mathcal{W}_{D_{peak}}$ is simply the product of two inter-scale saliency measure on both sides of the peak in entropy.

$$\mathcal{W}_{D_{peak}} = \mathcal{W}_D(\hat{s_s}, \mathbf{x})\ \mathcal{W}_D(\hat{s_s} + \Delta s, \mathbf{x}) \tag{3.8}$$

Multiplying the two inter-scale saliency terms serves only to maximise the separation between salient and less salient regions and also provides some measure for mutual support. Note that multiplying in this case does not provide a measure of correlation, which would favour symmetric peaks. However, it suppresses the saliency value from regions where the

corresponding entropy-scale peak is not particularly strong for example if one side of the peak has a very shallow almost non-existent peak. Perhaps a more intuitive method of quantifying the characteristics of the peak would be to use a central difference of the entropy values around the peak. However, this would not suppress less pronounced peaks as much as using a multiplication. The immediate effect of the new inter-scale measure can be seen using the previous examples shown in Figure 3.1. The saliency measure for the tree, the eye and the sky regions have changed from $(6.18, 4.27, 0.03)$ to $(6.94, 8.05, 0.00)$ respectively.

A more detailed experiment is shown in Figure 3.2 from the same picture shown in Figure 3.1(a) where Kadir and Brady's scale saliency algorithm and our modified version was applied to three foreground and three background regions. The first row shows an enlarged version of the regions of interest where the sampling kernel scans from the lowest scale to the limits of the region shown. Here, a red, thicker-lined circle shows the scale attributed to the highest saliency value. The green and white circles show the less salient peaks in entropy that were found. In the next row, the entropy-scale characteristic is shown. Again, the thicker red line represents the peak in entropy that corresponds to the most salient scale and the thinner green lines show the rest of the peaks in entropy that were found. The following row shows the variation of the inter-scale saliency measure $\mathcal{W}_D$ over the same set of scales. Over each $\mathcal{W}_D$ curve, the highest saliency value extracted from the region is shown. We can see already from this row that the saliency values do not accurately represent the saliency of the regions since the order from most salient to least salient regions are:Eye1;Eye2;Tree1;Tree2;Face;Sky. If we observe the next row down, this shows the inter-scale measure, $\mathcal{W}_D$, multiplied by its value at the neighbouring scale. This is similar to a squaring function, which enlarges the variation in $\mathcal{W}_D$ over scales. Looking at the saliency values above each graph in this row, the ordering of the different regions becomes more accurate;Eye1;Eye2;Face;Tree1;Tree2;Sky.

More comparative results using different images are shown in Figure 3.3. Here the leftmost column shows the original grey-scale images, the middle column shows the results using the original scale saliency algorithm and the rightmost column shows the results using the modified algorithm, using Equations 3.7-3.8. In the two columns showing the results, red circles show the top N most salient spatial regions where the size of the circle indicates the

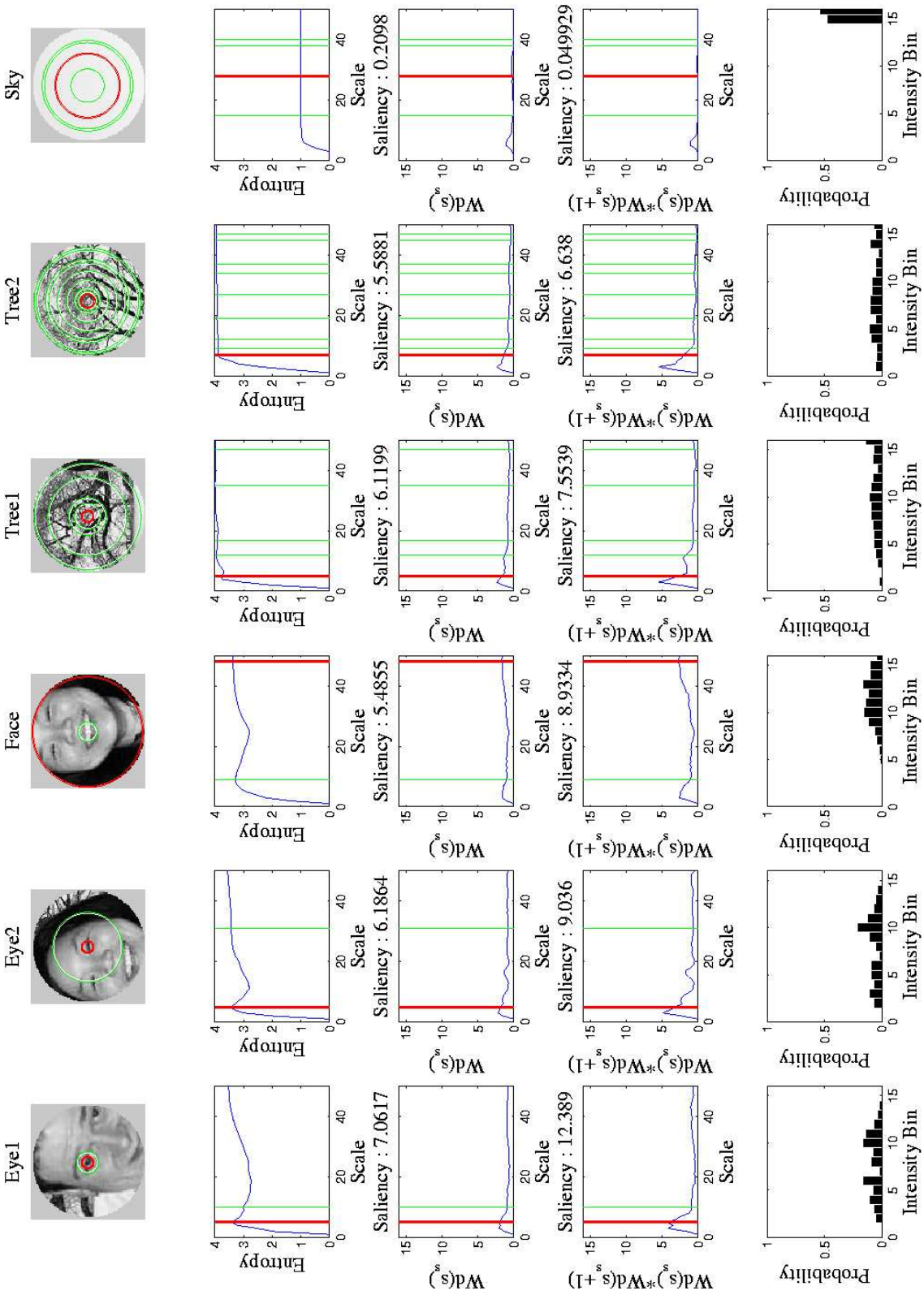


Figure 3.2: Entropy-scale, inter-scale saliency-scale curves and intensity histograms of 6 regions of foreground and background from the image in Figure 3.1(a).

Figure 3.3: Further comparison between Kadir and Brady's original method and the modification in Equation 3.7-3.8. The first column shows the original image, the middle column shows results using the original scale saliency algorithm and the rightmost column shows results using the modified version. The circles show the top N most spatially salient regions where the size of the circles indicates the salient spatial scale. The images used in the two result columns are shown in colour for easier discrimination between that and the salient spatial regions marked with red circles.

(a) (b)

Figure 3.4: The difference between the modified saliency algorithm and the original. Results are calculated on the grey-scale image in Figure 3.3(a). (a) shows the result using 50 of the top salient regions to highlight the same region while (b) only uses 15.

salient spatial scale.

In the first row of Figure 3.3, the result using our modified scale-saliency algorithm indicates a set of more tightly packed circles around the salient regions. This is likely to be due to the increased discriminance of our modified method. In the second row of Figure 3.3, although our modified version highlights more of the self-similar background regions, this is because the method is more discriminative and therefore, fewer circles around the same salient region yields very similar responses. Therefore, using fewer circles, in (f) it is possible obtain a very similar set of selected regions as (e), as shown in Figure 3.4 where (a) shows the original method, requiring more salient circles to highlight the same regions as the modified method, which requires only 15 regions to highlight the same region (beyond 15, it starts to highlight more and more regions).

Overall, the experiment above shows us that although the tree regions exhibit consistently high entropy at most scales, its saliency should still be relatively low since the texture is fairly self-similar. This notion of self-similarity in space can be extended to address problems of non-stationary background in time. Specifically, the inter-scale measure provides a description of the local variation in intensity distribution, which allows us to distinguish genuinely salient

homogeneous spatial regions and potentially spatio-temporal volumes from those that exhibit predictable behaviour over time.

Using a measure around both sides of the scale at which the entropy peaks provides a better indication of the overall shape of the entropy-scale characteristic. However, this still only measures the variation of the inter-scale saliency within the locality of the peak. Studying the entropy-scale characteristics in Figure 3.1, it is apparent that the tail of the curve (i.e. entropy values at higher scales) provides a descriptive measurement of the saliency of the overall spatial neighbourhood. Moreover, the entropy-scale characteristics of the tree regions are similar to each other but very different from those of the eye region. The purpose of extracting spatially salient regions, however, is not to identify exact similarities in the entropy curve but to provide some measure of the shape of it by measuring its variance. Therefore, the scale saliency algorithm is further extended to take into account the variance of the entropy,$\sigma_{\mathcal{H}_D(\mathbf{x})}$, over increasing scales.

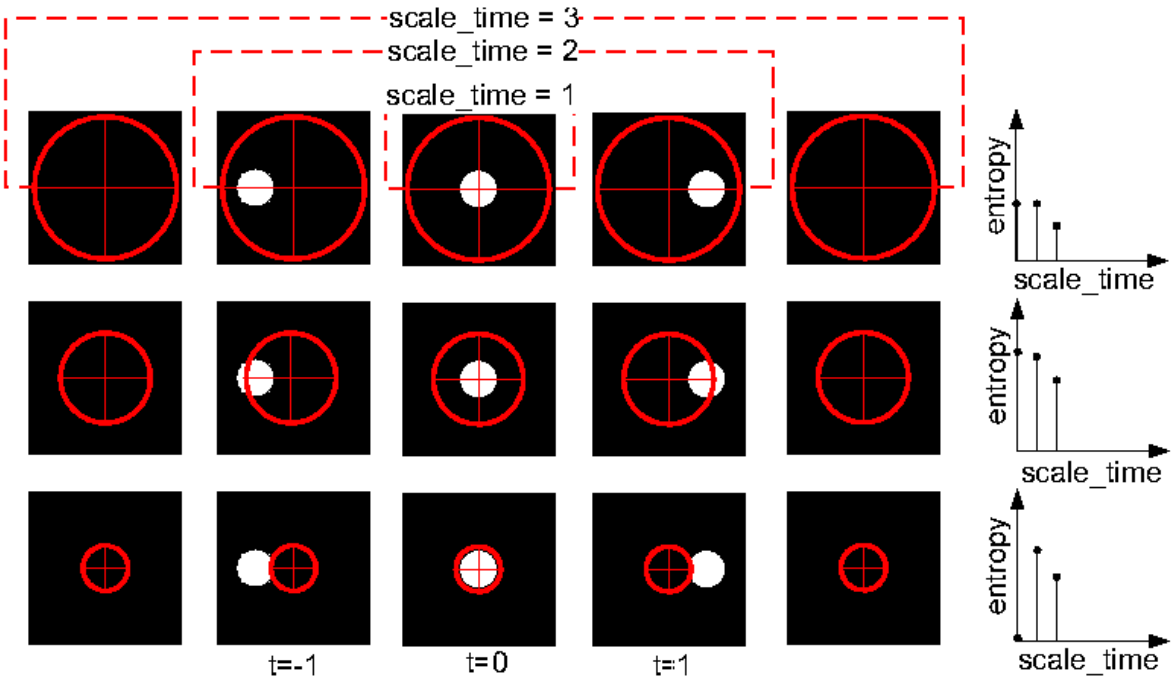
$$\mathcal{Y}_D(\hat{s}, \mathbf{x}) = \mathcal{H}_D(\hat{s}, \mathbf{x})\sigma_{\mathcal{H}_D(\mathbf{x})}\ \mathcal{W}_{D_{peak}} \tag{3.9}$$

where $\sigma_{\mathcal{H}_D(\mathbf{x})}$ is defined as:

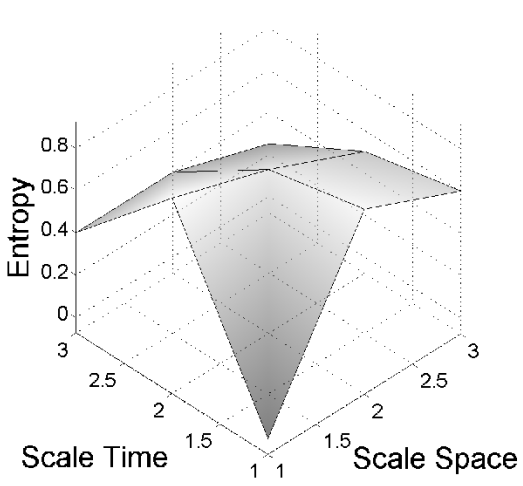
$$\sigma_{\mathcal{H}_D(\mathbf{x})} = \sqrt{\frac{1}{s_{s_{max}}} \sum_{s_s \in \{1, s_{s_{max}}\}} \left[\mathcal{H}_D(\mathbf{x}, s_s) - E[\mathcal{H}_D(\mathbf{x})]\right]^2} \tag{3.10}$$

3.3 From Spatial to Temporal Saliency

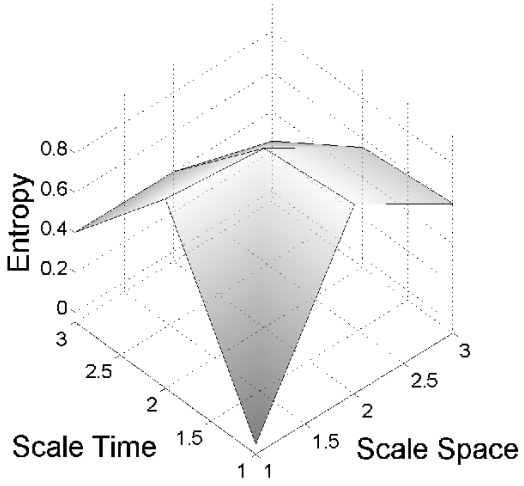
The previous section highlighted the potential of using a spatial saliency measure to separate cluttered background from foreground. To this end, the notion of saliency from a measure of spatial unpredictability is extended to temporal unpredictability. At its simplest form, something temporally unpredictable occurs when a particular spatial region appears or disappears over time from a sampled spatio-temporal neighbourhood. We can observe such behaviour by referring to Figure 3.5, which shows a synthesised sequence of a white dot moving from left to right on a black background. The manually calculated graphs in the rightmost column



(a)



(b)



(c)

Figure 3.5: Entropy-Scale characteristics of a dot passing over a plain background. (a) The same sequence of images with 3 different sizes of kernel. (b) The entropy scale characteristic centred at $t = -1$. (c) The entropy-scale characteristic at $t = 0$.

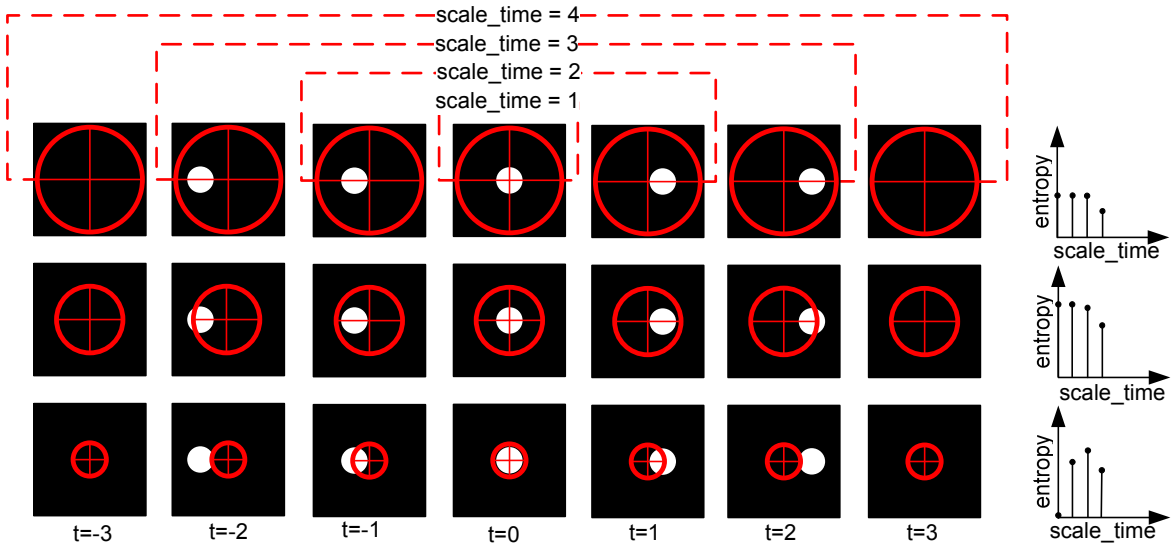
show the variation of entropy over temporal scales at a single spatial scale, centred at t=0. The two graphs (b) and (c) show the entropy-scale characteristics centred at time $t = -1$ and $t = 0$ respectively. The entropy-scale characteristic taken at $t = 0$ has a more pronounced peak at the middle scale in space and time than that evaluated at $t = -1$. This example illustrates that the entropy-scale characteristic is sensitive to temporal shift as well as spatial shift. If the dot moves slower across this kernel, as seen in Figure 3.6, then the peak in entropy is seen at a larger time scale. Therefore the method can also locate the temporal scale of salient motion.

The advantage of extending Kadir and Brady's scale saliency algorithm in this way is that self-similarity in space can be treated similarly over time. This can be illustrated with a simple example, shown in Figure 3.7 where the entropy-scale characteristics in space and time have been calculated for two contrasting areas of motion. These are shown in Figure 3.7(a) and (b), where enlarged frames of a pointing finger and region of moving hair respectively are illustrated. The circular region in each frame indicates the highest spatial scale at which the entropy-scale characteristic was calculated. The exact regions are highlighted by red circles in (d). Studying the entropy-scale characteristics in (c) and (e), which represent responses of the finger and hair regions respectively, we can see that (c) shows a distinctive peak in entropy, which has been marked by a red vertical line. In contrast, the entropy-scale characteristic in (e) shows very little variation over scale, thus highlighting the potential for using temporal saliency for discriminating between salient foreground and non-stationary background.

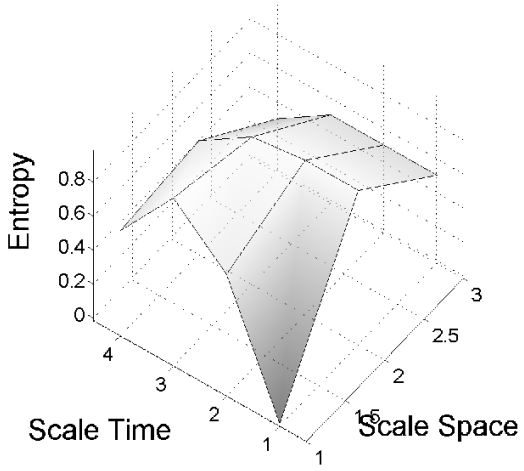
To calculate temporal saliency, we consider that the probability density functions of the intensity distribution are generated from spatio-temporal cylinders with a varying spatial radius s_s and temporal interval s_t. Therefore, the entropy of this cylindrical spatio-temporal volume around point $\mathbf{x} = [x\ y\ t]^\mathsf{T}$ is defined as:

$$\mathcal{H}_D(s_s, s_t, \mathbf{x}) = -\sum_{d \in D} b(d, s_s, s_t, \mathbf{x}) \log_2 b(d, s_s, s_t, \mathbf{x}) \tag{3.11}$$

where b is still the PDF of the intensity distribution and d is one of a set of D possible bin intervals of b, and s_s and s_t represent the spatial and temporal scale of the sampling kernel



(a)



(b)

Figure 3.6: Entropy-Scale characteristics of a dot passing over a plain background at a slower speed than in Figure 3.5. (a) The same sequence of images with 3 different sizes of kernel. (b) The entropy scale characteristic centred at $t = 0$ for 4 time scales and 3 spatial scales.

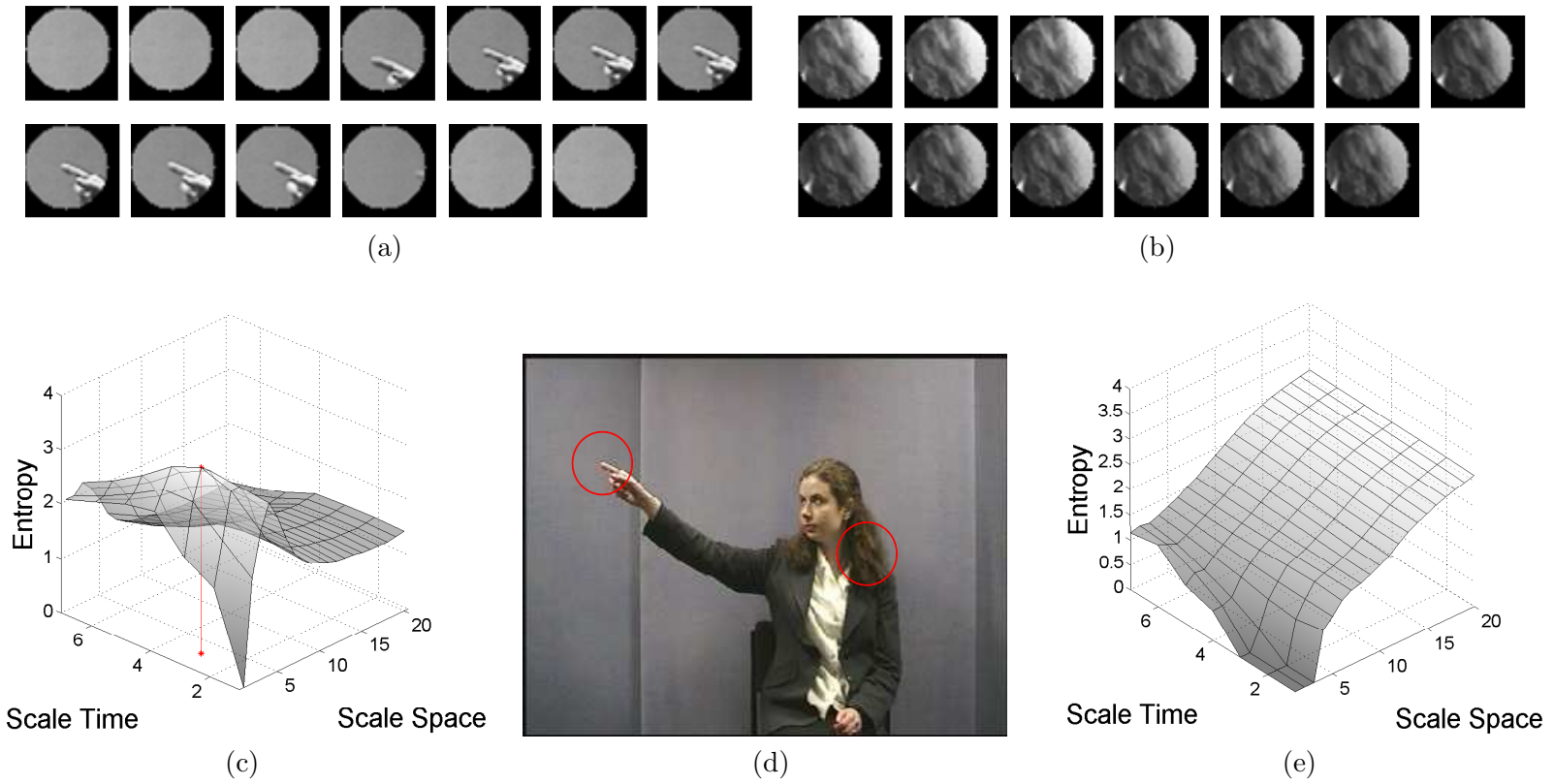


Figure 3.7: Study of applying Equation 3.13 to salient motion and self-similar motion. (a) and (b) show changes in the spatial window over all the temporal scales that were considered for the hand and hair region identified in (d). (c) and (e) show the corresponding entropy-scale characteristic for the hand and hair region respectively. A peak in entropy over spatial and temporal scale is identified by a red line in (c).

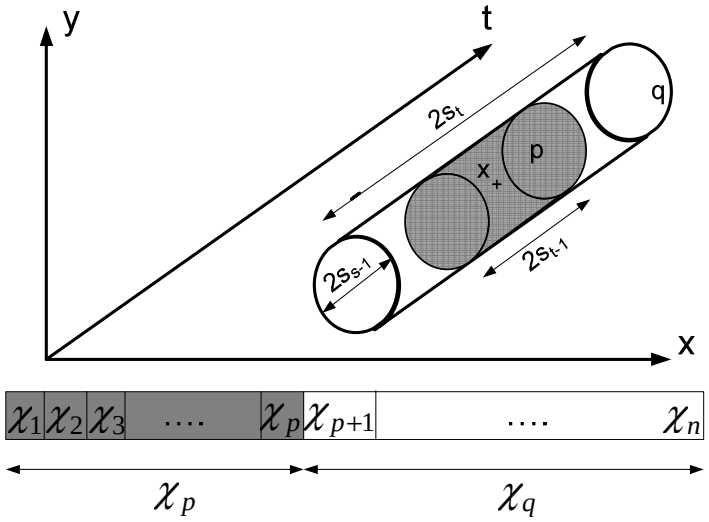


Figure 3.8: Inter-scale volumes. The pixels in the smaller cylindrical volume, shaded in grey contain the pixels in set χ_p and the pixels in the remaining white volume contain the pixels in the set χ_q such that the whole set of the pixels in the entire cylindrical volume is represented by $\chi_q \cup \chi_p$.

respectively.

The inter-scale saliency measure $\mathcal{W}_D$, which was defined in Equation 3.3 becomes a two dimensional matrix, representing the differential of the entropy for varying scales in space and over time. A normalisation factor is calculated for cylindrical volumes rather than circles. Since the proportionate increase in volume as the temporal scale is increased is not constant, a normalisation term is needed. This can be derived by first considering the cylindrical volumes shown in Figure 3.8 and how they contain two different sets of pixels. Here, the set of pixels $\chi_q \cup \chi_p$ contains all the pixels in the cylindrical volume defined by the scales $\{s_s, s_t\}$ and the set χ_p represents those within the smaller temporal scale $\{s_s, s_t - 1\}$. Hence inter-temporal scale saliency is defined as:

$$\mathcal{W}_D(s_s, s_t, \mathbf{x}) = s_t \sum_{d \in D} |b(d, s_s, s_t, \mathbf{x}) - b(d, s_s, s_t - 1, \mathbf{x})| \tag{3.12}$$

where s_t is the inter-scale normalisation factor for the cylindrical volumes shown in Figure 3.8, b is the PDF of the intensity consisting of D bins, s_s is the spatial scale. Taking into account both sides of the *temporal* peak in entropy, $\mathcal{Y}_\mathcal{D}$ is again defined as the scalar product

of the inter-scale saliency measure and its entropy value evaluated at $\hat{s} = [s_s \ s_t]^{\mathsf{T}}$ at which the spatio-temporal entropy peaks:

$$\mathcal{Y}_D(\hat{s}, \mathbf{x}) = \mathcal{H}_D(\hat{s}, \mathbf{x}) \ \mathcal{W}_{D_{peak}} \tag{3.13}$$

where the inter-scale saliency is redefined in terms of the spatio-temporal scale at which the entropy peaks.

$$\mathcal{W}_{D_{peak}}(\hat{s}) = \mathcal{W}_D(\hat{s}, \mathbf{x}) \ \mathcal{W}_D(\hat{s}+\Delta s_t, \mathbf{x}) \tag{3.14}$$

and the peak $\hat{s}$ is defined as the spatio-temporal scale at which the entropy peaks over the 2D entropy-scale surface.

$$\hat{s} = \{[s_s \ s_t]^{\mathsf{T}} : \hat{s}_s \wedge \hat{s}_t \wedge \hat{s}_{st}\} \tag{3.15}$$

and

$$\begin{aligned}
\hat{s}_s &= \{s_s : (\mathcal{H}_D(s_s-\Delta s_s, s_t, \mathbf{x}) < \mathcal{H}_D(s_s, s_t, \mathbf{x})) \wedge (\mathcal{H}_D(s_s, s_t, \mathbf{x}) > \mathcal{H}_D(s_s+\Delta s_s, s_t, \mathbf{x}))\} \\
\hat{s}_t &= \{s_s : (\mathcal{H}_D(s_s, s_t-\Delta s_t, \mathbf{x}) < \mathcal{H}_D(s_s, s_t, \mathbf{x})) \wedge (\mathcal{H}_D(s_s, s_t, \mathbf{x}) > \mathcal{H}_D(s_s, s_t+\Delta s_t, \mathbf{x}))\} \\
\hat{s}_{st} &= \{s_s : (\mathcal{H}_D(s_s-\Delta s_s, s_t-\Delta s_t, \mathbf{x}) < \mathcal{H}_D(s_s, s_t, \mathbf{x})) \wedge (\mathcal{H}_D(s_s, s_t, \mathbf{x}) > \mathcal{H}_D(s_s+\Delta s_s, s_t+\Delta s_t, \mathbf{x}))\}
\end{aligned} \tag{3.16}$$

where the scalar values $\hat{s}_s$, $\hat{s}_t$, $\hat{s}_{st}$, describes the scale at which a peak in spatial, temporal, and spatio-temporal entropy respectively is found. It is important to highlight here that the method for calculating the inter-scale saliency is differs from that proposed by (Oikonomopoulos et al., 2005) who suggested that changes in the intensity histogram over spatial or temporal scale were inter-changeable. It is also worth noting here that the peak in the spatio-temporal entropy-scale surface is not finding a peak in all 4 possible directions since it does not make sense to compare the inter-scale saliency over consecutive scales if s_t increases at the same time as s_s decreasing or vice versa. That is, the pixels in the cylindrical kernel in the previous scale are not a subset of those in the kernel at next spatio-temporal scale combination.

Each pixel location is represented by a feature vector, containing the location, spatial and temporal scales, and saliency found centred at that point. All selected features in each frame are ranked in order of saliency. Features are selected based on a threshold saliency value.

The Temporal Saliency Algorithm

1. The image is divided uniformly into a grid. The entropy is calculated at each grid location.

2. The entropy-scale characteristics,($\mathcal{H}_D$), and inter-scale saliency, ($\mathcal{W}_D$), are calculated using Equations 3.11 and 3.14 respectively. Each calculation is made centred at each grid location and time frame for all spatio-temporal scales in the interval $\{\{s_s, s_t\} : \{1,1\} \leq \{s_s, s_t\} \leq \{s_{s_{max}}, s_{t_{max}}\}\}$.

3. A saliency measure is calculated based at the location of the peaks in entropy over temporal, spatial and spatio-temporal scales using Equation 3.13

4. After the saliency has been calculated for all frames of the sequence, the peaks in entropy are ranked according to their corresponding saliency value. The total number of frames in the sequence must exceed the length of the kernel in temporal space.

5. The most salient time intervals and spatial regions are located according to a threshold on the ranked salient spatio-temporal locations.

3.4 Experiment

The temporal saliency algorithm was implemented using C++ and experiments were run on video data from 4 different outdoor and indoor scenes. Typical frames from the scenes are shown in Figure 4.5. In each of the sequences, different types of activity exist. In some cases, there are a mixture of people and cars, leading to large differences in potentially salient motion, and in other cases, there are sudden lighting changes either from reflections or from global changes of the ambient light. The length of the videos that was computed over varied according to the level of activity in each sequence.

To begin with let us analyse in detail a sequence from an outdoor scene that is highly cluttered and under constant change, as shown in row (a) of Figure 3.9. The pitfalls of analysing this scene is that there are rapid variations in lighting and the windy conditions that cause a bush at the bottom of each frame to be prone to a lot of motion. The wind also causes the camera to move on several occasions, making the entire scene shift slightly.

Figure 3.9: Example frames from 4 video sequences that were used. (a) shows a busy forecourt, (b) shows an indoor simulated shop scenario, (c) shows a busy motorway scene, and (d) shows an indoor scene where the camera faces out of a shop window.

Furthermore, the scene is taken at a very low frame rate so that methods for measuring direction of motion such as (Freeman and Adelson, 1991) would be impossible since there is no guarantee that there is any spatial overlap of the moving object between frames.

At each frame, the temporal saliency was calculated over a 7×10 grid to reduce computation time. For each grid location, the entropy-scale characteristic was calculated. Saliency values were calculated for any spatio-temporal peaks that were found where the maximum number of spatial and temporal scales $\{s_{s_{max}}, s_{t_{max}}\}$ was $\{20, 70\}$, and the grid size was 31 pixels vertically and horizontally where the image frame is 240×320 pixels. In order to observe the results of our temporal saliency algorithm, any temporally salient cylindrical volumes are superimposed over the relevant part of the sequence in both space and time. Therefore, if the salient spatial and temporal scale at a particular location was 10 and 4 respectively, this means that a circle of radius 10 pixels would be drawn within the time interval spanning 4 frames previous to and after the current frame. Therefore, the spatial size of the circle indicates its salient spatial scale and its persistence over time indicates its temporal scale.

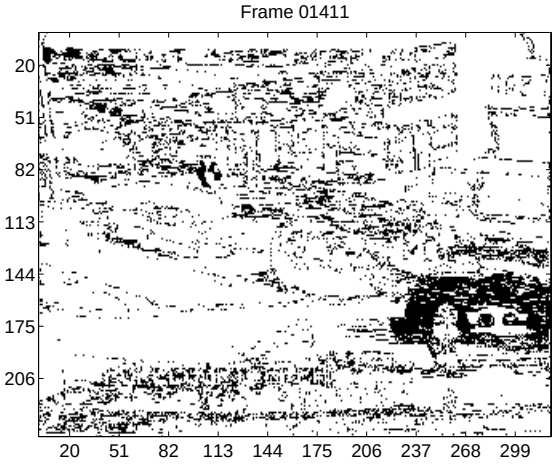
Our temporal saliency method performs significant background suppression, as shown by the comparative results using thresholded temporal frame differencing and spatial saliency in Figure 3.10. In (a) the thresholded temporal difference of a particular frame is shown where the black regions represent regions of higher intensity difference between frames. Visually, we can see that the temporal differencing identifies the moving car, movement from the bush, and reflections from the building at the far end of the image frame. In Figure 3.10(b), the results using spatial saliency are shown for the top 50% of most salient regions. Here the spatial saliency was similarly calculated over the same sub-sampled grid of the image to enable a fair comparison. However, a higher percentage of salient spatial areas are shown since there were much fewer saliency calculations that were extracted using the spatial saliency algorithm. The white circles represent the top 10% of most salient spatial regions while the remaining 40% is shown by black circles. The size of the circles indicates the salient spatial scale that was identified using Equation 3.5. The spatially salient circles are relatively evenly spread over the scene and under such severe sub-sampling, discrimination between salient spatially homogeneous regions is much more difficult.

Ideally, we would like to suppress the motion from trees and reflections as much as possible and this is achieved in Figure 3.10(c), where the results using our temporal saliency algorithm are shown. The top 1% of most saliency values are circles with a white line and the next 4% are indicated by black circles. Clearly much more distinctive regions of temporal and spatial saliency are shown. In particular, it is possible to identify the trajectory of motion of the car along the road as it has generated all the responses along its path. There are still a few salient regions that overlap the bush in the front of the scene, as well as some responses around spatial edges where sudden motion of the camera would cause spurious changes in the image intensity. Reflections from the glass windows at the back of the scene are also detected. However, compared to Figure 3.10(b), much less of the spatially salient parts of the scene are selected. Furthermore, the video was labeled with text at the bottom of the frame. This is highly salient when calculated with just spatial saliency but is not detected at all using temporal saliency.

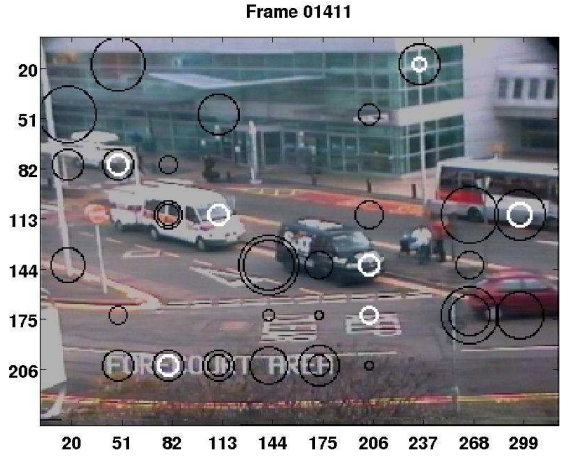
It is important to note that the heavy sub-sampling of the image frame has caused the spatial saliency algorithm to perform much less well than it might have done otherwise, since it largely relies on spatial features coinciding within the local spatial neighbourhood of the circular kernel. However, with temporal saliency, even under heavy sub-sampling, the method is more robust to spatial aliasing since the temporal saliency can be relied upon more to determine approximate spatial regions which have temporally salient behaviour. Note however, that due to the low frame rate of the sequence, it was very possible for an object to move 60 pixels horizontally[2], which explains the discontinuity of the responses along the path of the moving car in Figure 3.10(c).

It is evident that our temporal saliency algorithm does not pick out as much global motion as temporal frame differencing and hence provides a degree of background motion suppression. In particular, the noisy motion of the bush at the bottom of the scene had much lower saliency values, when compared to that of the moving car. On some rare occasions, parts of the bush were considered foreground due to sudden motion patterns caused by gusts of wind. Although this is not the ideal results from the algorithm, it is in keeping with the nature of

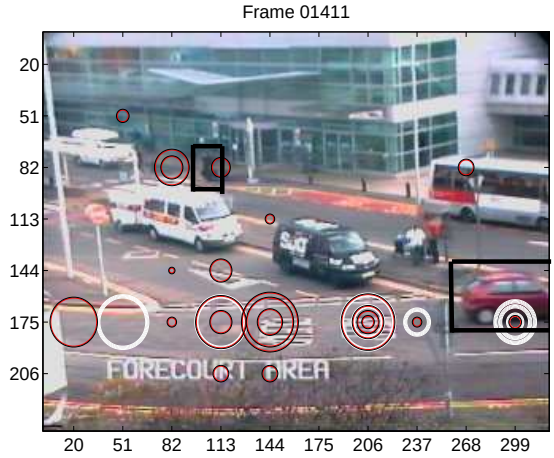
[2] which is greater than the resolution of the sub-sampled grid, which was 31 pixels apart



(a)



(b)



(c)

Figure 3.10: Detailed study of the results from a highly cluttered scene run over 60 frames with comparisons to temporal frame differencing and spatial saliency. (a) Consecutive frame temporal differencing using a threshold of 20. Higher differences are shown in black. (b) 10% most salient regions using the spatial saliency algorithm, and (c) the 5% most salient regions using the temporal saliency algorithm.

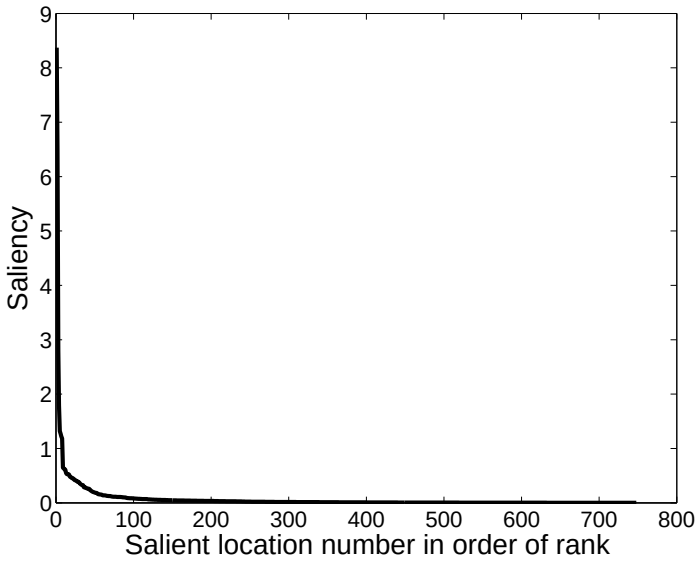


Figure 3.11: Graph showing ranked saliency values extracted from a typical frame from the sequence shown in row (a) of Figure 4.5. We can see that there is a clear separation between salient foreground and less salient background.

it. For more sophisticated background foreground separation, one would require higher levels of representation of the temporal dynamics of the scene. However, in general, the highest saliency value caused by the bush and moving car was 0.55 and 8.37 respectively. The strong background motion suppression is further demonstrated in Figure 3.11 where the saliency is plotted in order of rank for a typical frame of the sequence.

We can study the results of the busy forecourt sequence further in Figure 3.12. Here, the top graph shows the highest saliency value at each frame showing two clearly separable events, defined around the peaks {A-C} and {D-F}. The peaks that are indicated by a letter have their corresponding frame as indicated in the 3 bottom rows of the figure. The middle two rows show key frames around peaks {A-C} and the remaining row shows the key frames in the cluster of peaks {D-F}. Here, the top 5% of saliency values are indicated by circles where the top 1% are white circles and the remaining 4% are shown in black. For comparison against the ground truth, rectangular regions have been marked out in each of the frames to show the areas containing salient motion. If the circles do not overlap a spatially salient region, salient motion occurred either before or after the current frame. This is because the cylindrical kernel is spread over past and future frames as shown in Figure 3.5(a). Since

the circles appear symmetrically in time about the central frame it is not always possible to identify whether salient motion at a particular time scale occurred before or after this central frame. For clarity, where possible, only salient regions whose temporal peak occurred with an observable interval around the central frame are shown.

In each of Figures 3.13-3.15, three of the most salient frames for 3 different video sequences from both indoor and outdoor scenes are shown. In each case, meaningful salient motion, within the context of the chosen sequences, was detected. Again, the circles indicate regions of salient motion. The top graph in each of Figures 3.13-3.15 show the highest saliency value centred at each frame. The letter of each marked peak in saliency is shown in the bottom right hand corner of the corresponding frame. The white circles show the top 1% whilst the black shows the next 4%. The black boxes show manually selected salient regions. For clarity, of the top 5% most salient locations, only those which resulted in temporally salient peaks at temporal scales $s_t \leq 5$ are shown. Again, circles that do not appear directly in line with the salient moving object of the frame are caused by salient motion which occurs before or after that frame. From inspection of the frames, it is possible to interpolate the salient motion of the objects.

Figure 3.13 shows the results from a busy traffic scene. Within this sequence, many cars move along the roads. However, at a certain point, a car stops and reverses into oncoming traffic. The grid size was set particularly small at 11 pixels in just this set of results in order to accommodate for the large amounts of diagonal motion in the scene. In this scene, the frame rate was much higher so it is easier to identify areas of motion with spatially salient overlapping regions. The same has a lot of noisy motion caused by the traffic, which the algorithm finds difficult to separate from more unusual motion, such as the reversing car. This particular event was still amongst the more salient time intervals detected and is highlighted at peak B of Figure 3.13. However, the saliency value of this particular event is not the highest for the saliency measures calculated, centred at this frame.

Figure 3.14 shows the results from a simulated scene of a drinks shop where there is a shopkeeper seated to the right of the scene and customers enter from the left to browse the selection of drinks cans on the left side. The scene is highly cluttered and the shopkeeper to

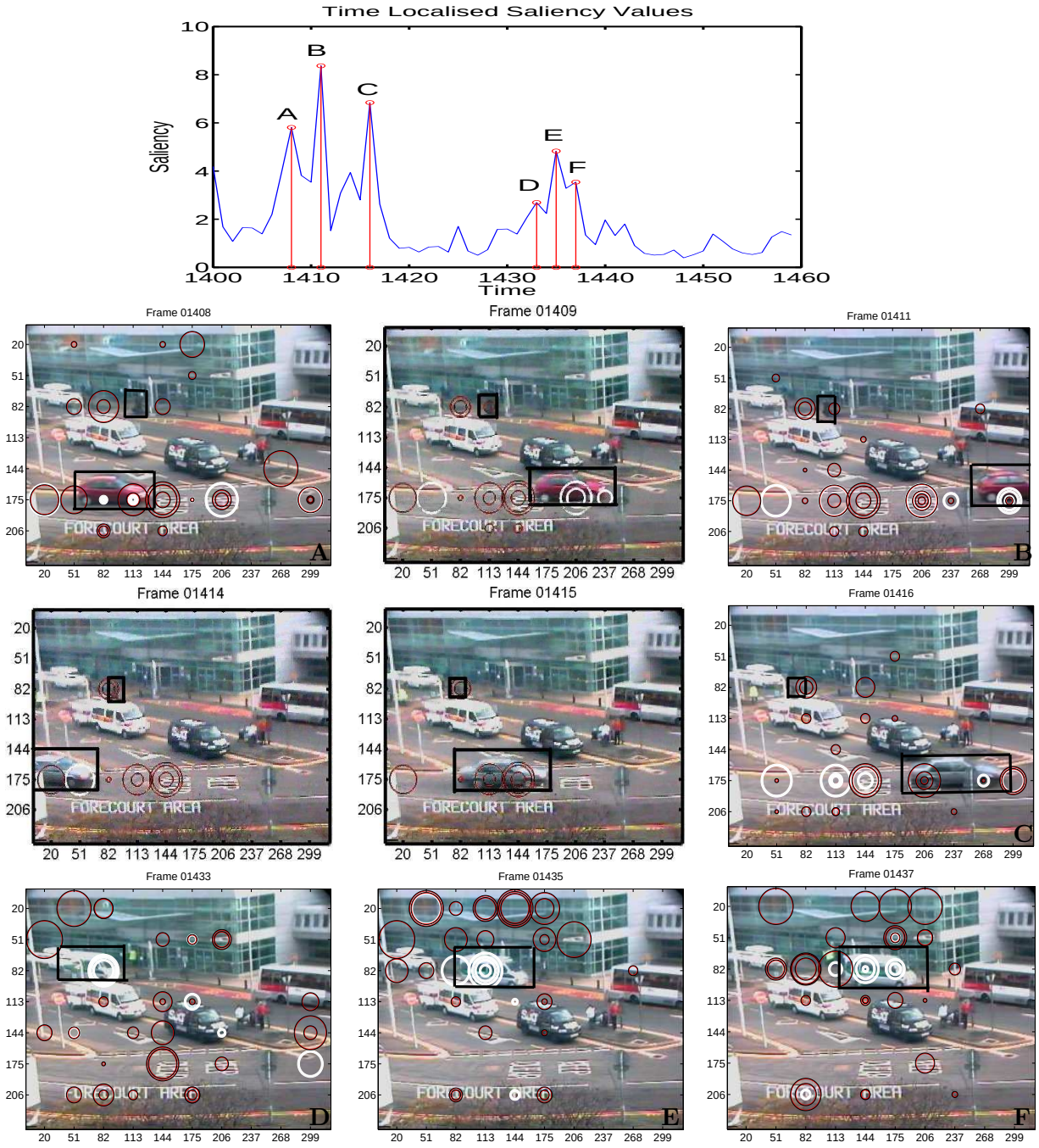


Figure 3.12: Results from a highly cluttered scene run over 60 frames. The top graph shows the highest saliency values at each frame plotted over time. Peaks that have been indicated by a letter correspond to the indicated frames in the rest of the figure. The bottom 3 rows show key frames with circles indicating the areas with high temporal saliency. The size of the circles indicates the spatial scale at which the temporally salient volume was identified. The top 1% most temporally salient regions are shown in white and the following 4% are shown in black. The first two rows of image frames show key frames around the peaks {A-C} and the bottom row shows the key frames around peaks {D-F}. Black rectangles are manually highlighted to show parts of the scene that exhibit temporally salient motion.

the right of the frame causes some noisy background motion that is not considered salient compared to the motion of the customer. The results clearly mark out two different customers and the graph at the top of Figure 3.14 indicates time-localised intervals of salient motion.

Figure 3.15 shows a scene taken through a shop window. Hence the sequence has many light reflections, and sudden changes in light caused by the poor quality of the video capture. The likely areas of interest are towards the back of the scene, behind the glass when passers by browse at the shop window. The experiment was run over a relatively short sequence and as a result, time-localised peaks in the graph of column (c) are not apparent. The circles represent salient motion detected over a maximum of 21 scales. The results indicate different stages of activity for two pedestrians walking past the window. There are some anomalous circles which were caused by the sudden change in lighting though, these locations tended to be less salient.

3.5 Discussion and Conclusions

The method described in this chapter shows an interesting and powerful method for selecting and representing temporal saliency in busy natural indoor and outdoor scenes. In particular, the algorithm demonstrates the power of using contextual information for representing and extracting low and mid level features. The results shown in the previous section show that relying only on the scene data may still yield quantifiable differences between a mixed collection of activities.

A problem with automatically interpreting real scenes is that it is very difficult to separate noisy background clutter from salient motion. In these results, noisy background clutter from non-stationary regions such as foliage, as well as small motions from a moving camera, subtle lighting changes, and reflections could be suppressed.

The disadvantages of the temporal saliency algorithm is that it is not possible to identify whether salient motion occurs before or after the central frame at which it is detected, though this was rectified by applying a one-sided sampling kernel. Furthermore, the algorithm is computationally expensive since it computes a new histogram at every location for every

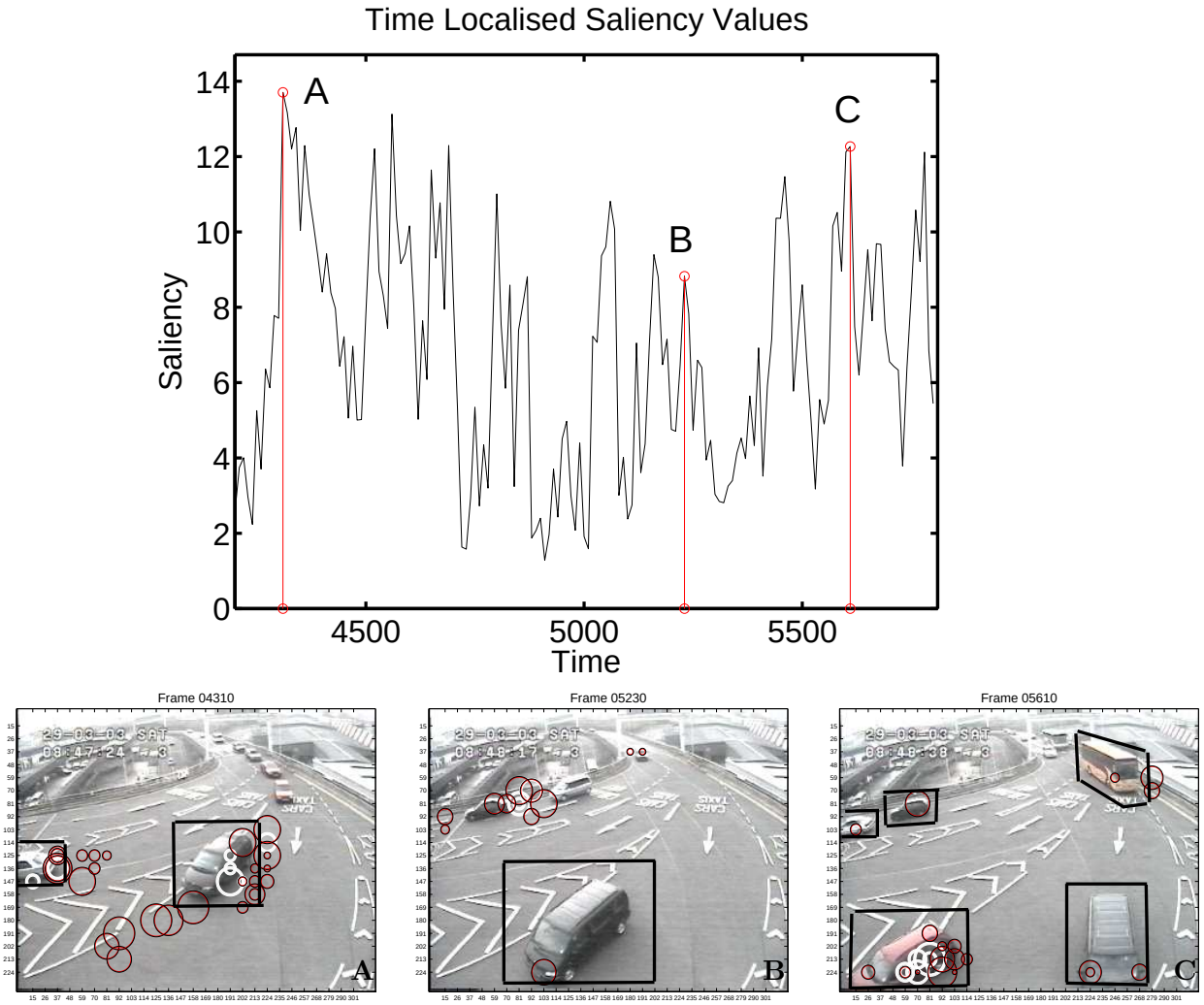


Figure 3.13: Results from an outdoor scene of a busy road run over 1400 frames. The letter of each marked peak in saliency is shown in the bottom right hand corner of the corresponding frame. The white circles show the top 1% whilst the black shows next 4%. The tick marks on each frame indicate the locations around which the temporal saliency algorithm was calculated.

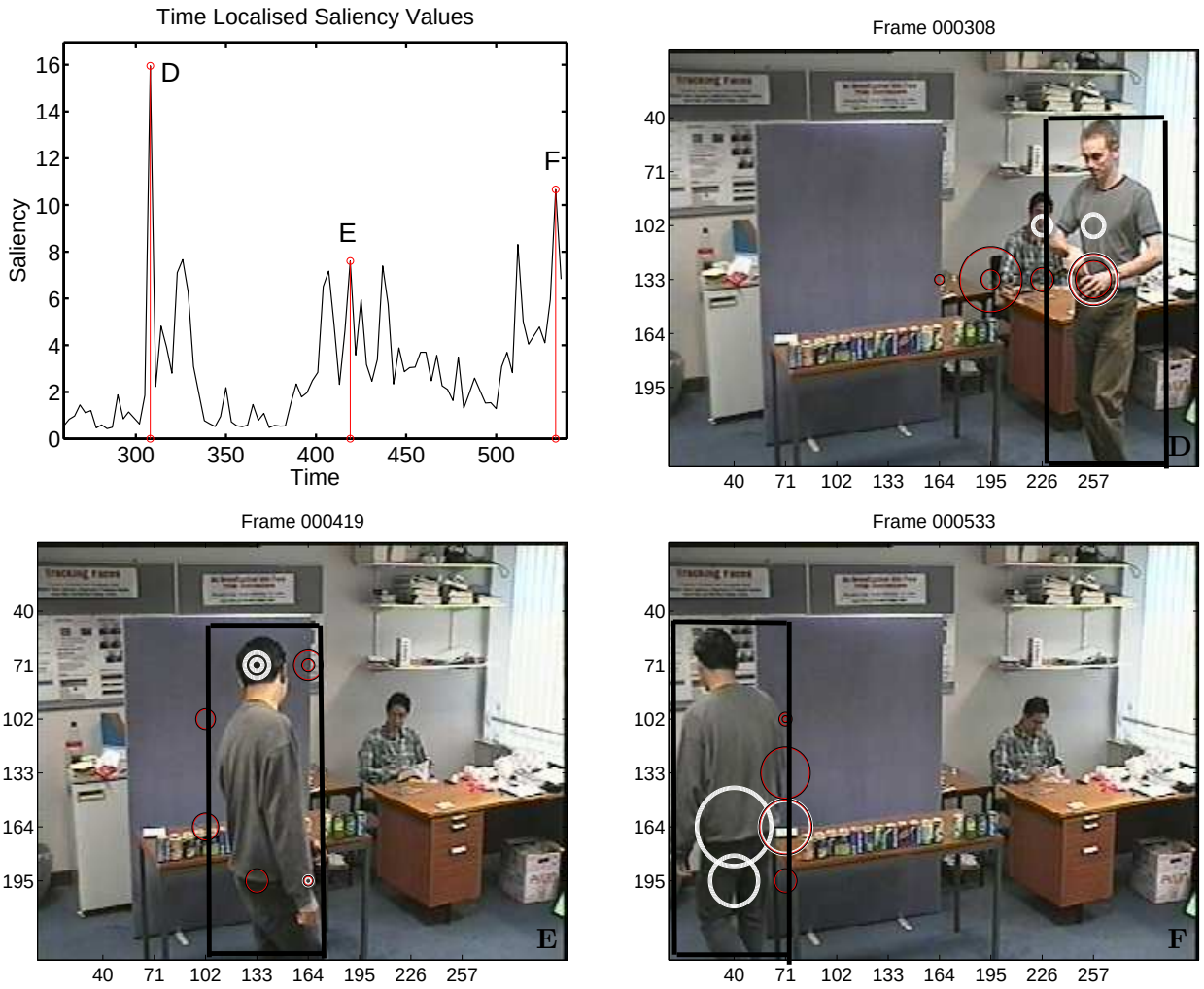


Figure 3.14: (b) Simulation of a drinks shop run over 372 frames . The letter of each marked peak in saliency is shown in the bottom right hand corner of the corresponding frame. The white circles show the top 1% whilst the black shows next 4%. The tick marks on each frame indicate the locations around which the temporal saliency algorithm was calculated.

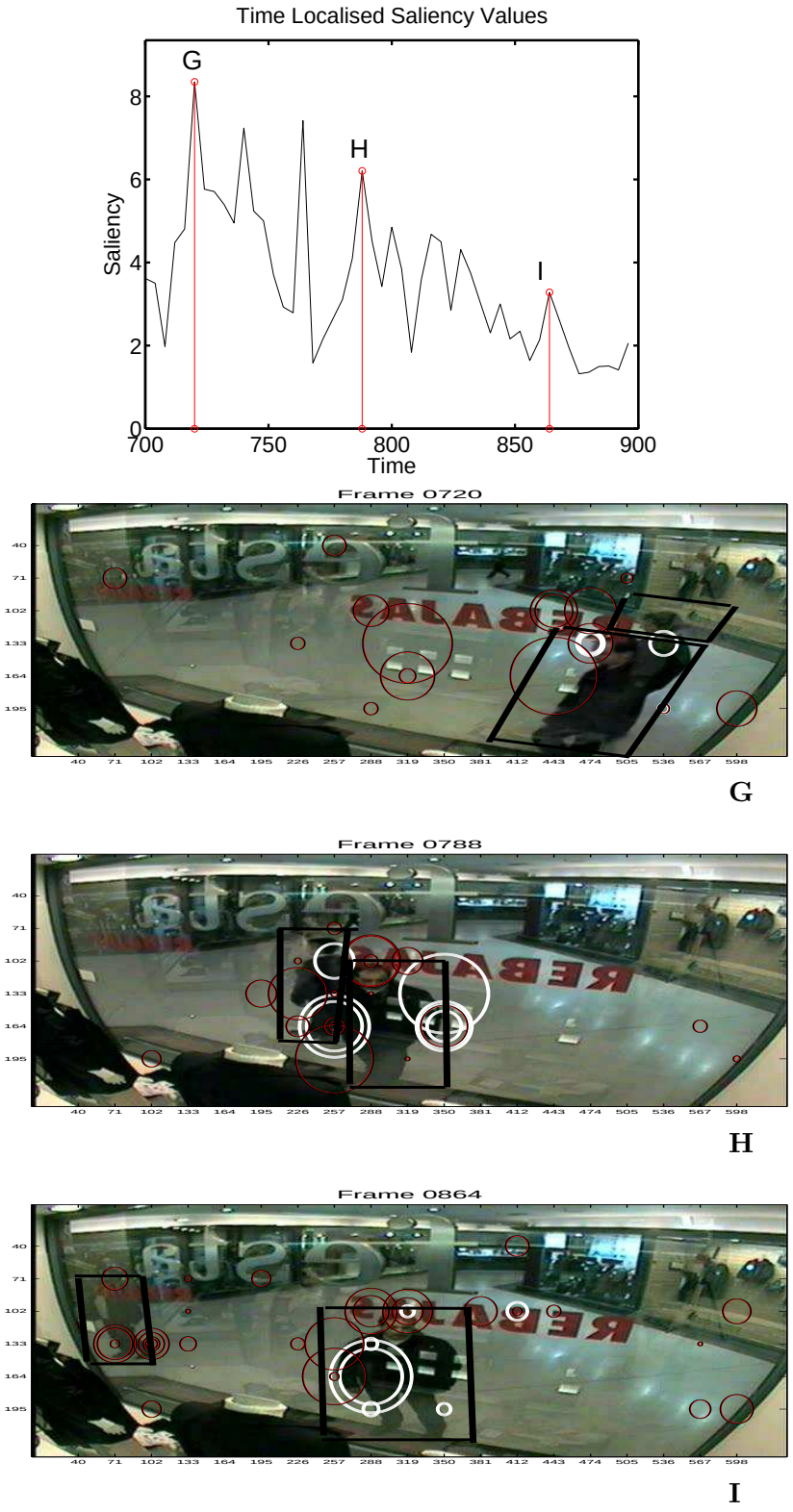


Figure 3.15: Results from cluttered scene through a shop window run over 150 frames. The letter of each marked peak in saliency is shown in the bottom right hand corner of the corresponding frame. The white circles show the top 1% whilst the black shows next 4%. The tick marks on each frame indicate the locations around which the temporal saliency algorithm was calculated.

Figure 3.16: Examples of rarity rather than saliency in two frames of the busy forecourt sequence. The two frames highlight the distinct difference between the window of the bus in the centre of the frame when it catches the light. If this does not occur frequently in the video sequence, then this would yield an extremely high temporal saliency measure.

scale. It will be possible to reduce computational complexity by concentrating computation on more localised areas that are deemed to be more salient.

As yet, the algorithm is not able to distinguish between rarity and saliency. In one particular experiment, the most salient feature of a busy traffic scene was the light reflected off a vehicle (see Figure 3.16). Such anomalies would be ranked out of significance if the algorithm was run over a longer time interval, when enough re-occurrences of such anomalies were found. Another issue here was how the salient regions are selected once calculated. In this case, all saliency values in the whole sequence were ranked in order to find those that were most salient. However, using this method may mean that a salient region within a larger neighbourhood might not be considered as salient even though it is so within a mid-level local context. So for example, if neighbouring regions of interest are all salient to a similar degree, does this make the region salient any longer? This problem will be addressed in Chapter 5 where the temporal saliency algorithm is modified to extract salient spatio-temporal regions from a busy outdoor scene.

Another disadvantage of the algorithm is that there is an inherent bias of the algorithm to smaller spatial scales where the likelihood of salient variations in the local spatio-temporal pixel intensity are increased. It is possible to rectify this to some extent, by applying the

spatial saliency algorithm first to find seed locations before performing temporal saliency, which will be shown in Chapter 5. Investigations will also be made into quantifying spatio-temporal saliency in Chapter 5 where the difference between temporal saliency and spatio-temporal saliency is subtle but important for the feature selection process.

4 Correlating salient interactions

The previous chapter approached the problem of feature extraction and representation. However, on their own, temporally salient regions can only indicate regions of interest but do not, individually infer higher semantic levels of behaviour. We also saw that it was possible to detect motion but there was no way to distinguish between different types of behaviour such as detecting a car stopping and reversing into oncoming traffic. In this chapter, the property of feature binding is investigated by finding co-occurrences between spatio-temporally proximate as well as temporally proximate but spatially separated motion.

It is desirable to find ways to bind and model temporally correlated behaviour since we can use this to accumulate usual patterns of activity and hence identify salient cause and effect phenomena from a scene. Intuitively, we can cluster low level features in order to track objects and understand typical scene topology. To this end, a popular approach is to cluster features and then perform co-occurrence analysis. Stauffer and Grimson (2000) modeled the background using Gaussian mixture models in order to identify unusual foreground pixels. Once these were bound together using spatial binding, object tracking was used to accumulate a model of usual motion trajectories. However, in highly cluttered scenes, where partial or total occlusion of an object occurs, it is not always viable to perform multiple object tracking (Xiang and Gong, 2006).

Catering for robust tracking under partial or total occlusion requires pre-determined contextual assumptions about what the scene may contain. Even fitting well-defined object models is not sufficiently robust to distinguish between moving objects in crowded scenes. That is, tracking objects is never perfect due to occlusion and changing lighting conditions. Many methods use complex rules to try and overcome problems due to occlusion (Khan et al.,

2003; Oh et al., 2004; Zhao and Nevatia, 2003) but these can never be accurate and degrade in particularly crowded situations. For example, Zhao and Nevatia (2003) approached the problem of tracking people in crowded scenes by using a Markov Chain Monte Carlo-based method for finding the Maximum a posteriori estimate of the likelihood of the whole image given the objects and the background model. Even with a 3-D representation of the human body shape with 3 ellipses, representing the head, torso and legs of each person, and also some knowledge of the ground plane, the model still lost track of some of the people in the crowded scene. This is particularly unhelpful in cases where we wish to classify the overall behaviour of a person (and perhaps their interactions with others) if the same person cannot be tracked reliably. Furthermore, in this example, the camera angle was such that the scene was taken from a 40° elevated view so that inter-person occlusion could be minimised. It will be shown later that it is possible to express the underlying patterns of motion from a sequence and leave tracking to higher-level contextually explicit scene understanding tasks.

This chapter addresses the problem of correlating salient motion both at a spatio-temporal level and across spatially separated regions since it is in the interactions that more meaningful scene interpretation can be found. We will show that it is possible to spatio-temporally locate and detect salient motion events and interactions in 2 contrasting scenarios using a single hierarchical co-occurrence framework. Again, all the results will be generated from the data alone, with no prior knowledge of what each scene contains.

The method accumulates co-occurrences of atomic salient motion descriptors based on spatio-temporally interacting neighbouring grid responses. That is unusual or salient motion caused by an individual event, or more complex multiple cause-effect phenomenon from spatially separated but temporally correlated scene locations will be detected. Specifically, salient events caused by motion that would not be considered particularly meaningful individually, may define a more sophisticated level of understanding when addressed in combination. The results at the end of this chapter will show that this information is inherently measureable from raw image sequence data and should not require external top-down contextual models.

The rest of this chapter describes how an initial layer of co-occurrence on spatially close regions is calculated. A description of a higher level co-occurrence process will then be given

to find spatially separated but temporally correlated patterns of motion. Experimental results will demonstrate that it is possible to bypass issues of model order selection and complexity at the initial stages of feature extraction which tend to involve clustering noisy and arbitrarily thresholded functional responses from the imagery data (Zhong et al., 2004; Stauffer and Grimson, 2000).

4.1 Spatio-temporal binding

The disadvantage of the temporal saliency descriptor in the previous chapter is that the measure is completely invariant to motion direction and this is detrimental for differentiating between different types of salient motion. Accumulating spatio-temporal co-occurrences of the temporally salient features would solve this problem partially since a model of the most likely temporal saliency values within a local temporal neighbourhood is constructed. A more descriptive model, would also be able to identify the most likely configuration of saliency values within a local spatio-temporally connected neighbourhood.

Identifying likely spatio-temporally connected co-occurrences of saliency values addresses the problem of finding a representation for likely directions of motion without explicitly applying image patch correlations over time such as for methods like optical flow. Furthermore, by correlating saliency values, a more generic form of spatio-temporal binding is possible since the configuration of temporal saliency responses of a particular region can be approximated to a locally discriminative descriptor for a particular moving object.

Since a co-occurrence matrix can be considered a histogram, let us first define a 1-dimensional histogram b, which describes the probability distribution of data set $\mathbf{A}$ where its i^{th} bin is defined as

$$b_{\mathrm{i}} = \sum_{q=1}^{|\mathbf{A}|} \delta[bin(a_q) - \mathrm{i}] \tag{4.1}$$

where $bin(a_{\mathrm{q}})$ defines the index of the bin associated with the q^{th} value of data set $\mathbf{A}$, $|\mathbf{A}|$ defines the cardinality of the set, i ranges from 1 to M bins, and $\delta[\cdot]$ is the delta function which is equal to 1 when $bin(a_q) - \mathrm{i} = 0$. The bin index for each value in $\mathbf{A}$ is $bin(a_{\mathrm{q}})$, which

by definition is an integer and can be extracted by the relation:

$$bin(a_{\mathrm{q}}) = (a_{\mathrm{q}} - min(\mathbf{A}))\frac{M}{max(\mathbf{A}) - min(\mathbf{A})}, \text{ where } bin(a_{\mathrm{q}}) \in \mathbb{I} \tag{4.2}$$

Here, $max(\mathbf{A})$ and $min(\mathbf{A})$ define the maximum and minimum value in data set $\mathbf{A}$. A 2-dimensional co-occurrence matrix or histogram b of dimensions $M \times M$ given a set of data $\mathbf{A}$ is similarly defined as the frequency of pairwise co-occurrences, of all possible combinations of data points, a_{q} and a_{r} in $\mathbf{A}$. Therefore the frequency of co-occurrence of values a_{q} and a_{r} in bin i and j is defined as

$$b_{\mathrm{ij}} = \sum_{\mathrm{q}=1}^{|\mathbf{A}|}\sum_{\mathrm{r}=1}^{|\mathbf{A}|}\delta[bin(a_{\mathrm{q}}) - \mathrm{i}]\delta[bin(a_{\mathrm{r}}) - \mathrm{j}] \tag{4.3}$$

where both i and j exist in one of M equally distributed intervals or a set of classes and $bin(a_{\mathrm{q}})$ and $bin(a_{\mathrm{r}})$ again define the bin index for the values of a_{q} and a_{r} respectively.

In the more specific case of spatio-temporally binding salient regions of interest, co-occurrences are accumulated within a local spatio-temporal neighbourhood such that the data set $\mathbf{A}$ at each spatio-temporal location $\mathbf{x}$ becomes the set of saliency values at the 9-pixel neighbourhood of $\mathbf{x} = [x\ y\ t]^{\mathsf{T}}$ in the previous frame, as shown in Figure 4.1(a). Now a_{q} is a temporal saliency value attributed to a particular pixel location, $\mathbf{x}_{\mathrm{q}} = [x_{\mathrm{q}}\ y_{\mathrm{q}}\ t]^{\mathsf{T}}$ and a_{r} defines the corresponding temporal saliency values within the local spatio-temporal neighbourhood of $\mathbf{x}_{\mathrm{q}}$ and is defined as $\mathbf{x}_{\mathrm{r}} = [x_{\mathrm{r}}\ y_{\mathrm{r}}\ t-1]^{\mathsf{T}}$ in the previous frame. More specifically, $a_{\mathrm{q}} = \mathcal{Y}_D(\hat{s}, \mathbf{x}_{\mathrm{q}})$, $a_{\mathrm{r}} = \mathcal{Y}_D(\hat{s}, \mathbf{x}_{\mathrm{r}})$, $\{x_{\mathrm{r}} : x_{\mathrm{q}} - 1 \leq x_{\mathrm{r}} \leq x_{\mathrm{q}} + 1\}$, and $\{y_{\mathrm{r}} : y_{\mathrm{q}} - 1 \leq y_{\mathrm{r}} \leq y_{\mathrm{q}} + 1)\}$ where $\mathcal{Y}_D(\hat{s}, \mathbf{x})$ is the temporal saliency at location $\mathbf{x}$ and spatio-temporal scale $\hat{s}$ which is defined in Equation 3.13, Chapter 3. Now the maximum and minimum values of the data set are the extreme values of the temporal saliency measures, calculated over the whole video sequence that is being considered. Ordinarily, each spatial neighbourhood, centred around the same pixel might yield more than one peak in entropy. For simplicity here, we make the assumption that only the peak with the highest saliency is accumulated. So that an element of the

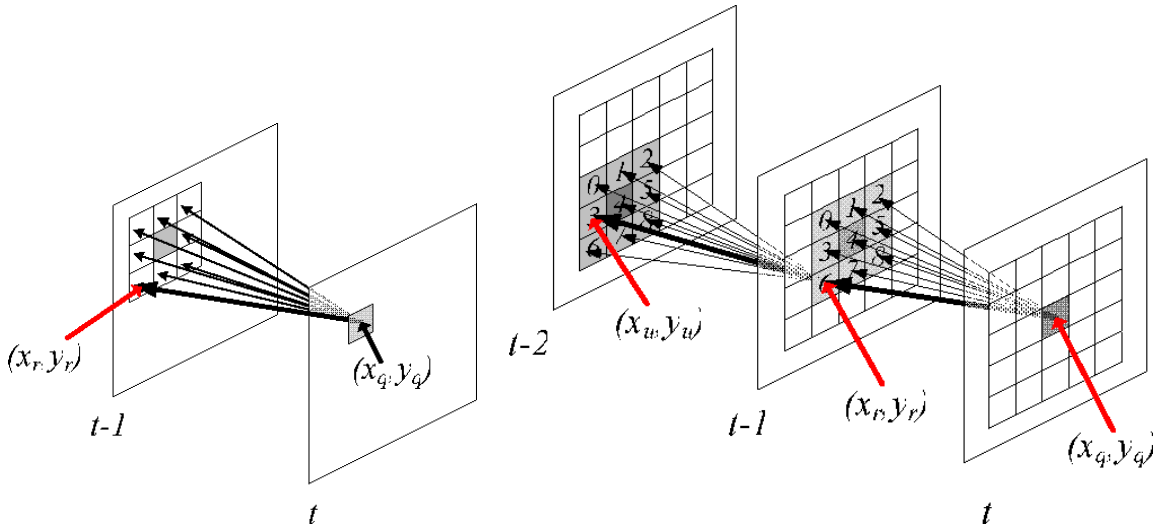


Figure 4.1: The co-occurrence of saliency values between frames. (a) A basic method of co-occurring saliency values based on the previous frame only. (b) Co-occurrence of saliency values over relative orientation based on previous 2 frames. Not all possible co-occurrences have been shown.

modified 2-dimensional co-occurrence histogram b is defined as

$$b_{\mathrm{ij}} = \sum_{\mathrm{q}} \sum_{\mathrm{r}} \delta[bin(\mathcal{Y}_D(\hat{s}, \mathbf{x}_{\mathrm{q}})) - \mathrm{i}] \delta[bin(\mathcal{Y}_D(\hat{s}, \mathbf{x}_{\mathrm{r}})) - \mathrm{j}] \tag{4.4}$$

where both q and r exist in the set of all possible locations for each frame and for every q, r only exists in the set of all pixels in the local neighbourhood of $\mathbf{x}_{\mathrm{q}}$ but in the previous frame. Here, $bin(a_{\mathrm{q}})$ and $bin(a_{\mathrm{r}})$ from Equation 4.3 have been replaced with their corresponding temporal saliency function for clarity. This provides a method of accumulating the most likely co-occurrences of temporal saliency values in a local spatio-temporal neighbourhood for each pixel location. However, the method could provide a more discriminative model of the local spatio-temporal saliency if the co-occurrence over the two previous frames was considered. In this way, the histogram becomes three-dimensional where each dimension refers to co-occurrences between a saliency value in the current frame and those in the previous frame within a local pixel neighbourhood, as shown in Figure 4.1(b). One element of the

co-occurrence histogram b for each pixel location $\mathbf{x}_{\mathrm{q}}$ is defined as

$$b_{\mathrm{ijk}} = \sum_{\mathrm{q}} \sum_{\mathrm{r}} \sum_{\mathrm{u}} \delta[bin(\mathcal{Y}_D(\hat{s}, \mathbf{x}_{\mathrm{q}})) - \mathrm{i}] \delta[bin(\mathcal{Y}_D(\hat{s}, \mathbf{x}_{\mathrm{r}})) - \mathrm{j}] \delta[bin(\mathcal{Y}_D(\hat{s}, \mathbf{x}_{\mathrm{u}})) - \mathrm{k}] \tag{4.5}$$

where q again can index any of the pixel locations in the current frame t, while r exists in the 9-pixel local neighbourhood of $\mathbf{x}_{\mathrm{q}}$ but in frame $t-1$ and u exists in the 9-pixel local neighbourhood around $\mathbf{x}_{\mathrm{r}}$ but at $t-2$. Therefore, $\mathbf{x}_u = [x_{\mathrm{u}}, y_{\mathrm{u}}, t-2]$, $\{x_{\mathrm{u}} : \{x_{\mathrm{r}} - 1 \leq x_{\mathrm{u}} \leq x_{\mathrm{r}} + 1\}$, $\{y_{\mathrm{u}} : y_{\mathrm{r}} - 1 \leq y_{\mathrm{u}} \leq y_{\mathrm{r}} + 1\}$ and $\mathbf{x}_{\mathrm{r}}$ is defined as before. b is accumulated over an entire frame of a sequence such that at time t, $b(t)$ is an accumulation of all possible spatio-temporal co-occurrences from the first to the current time.

If the the relative orientation between consecutive frames is also calculated, the co-occurrence histogram becomes even more descriptive. To achieve this, 2 extra dimensions are added to $b(t)$ by introducing the co-occurrence of relative orientation of a pixel location between consecutive frames as shown by the numbers in the highlighted pixels of Figure 4.1(b). Thus an approximate representation of possible temporally salient directions of motion is created if the scene is less busy. Or, if the scene is more cluttered, only a likely local spatio-temporal configuration of temporal saliency values is accumulated. Therefore, Equation 4.5 is modified to define b as a 5-dimensional co-occurrence histogram.

$$\begin{aligned} b_{\mathrm{c}} \quad = \quad & \sum_{\mathrm{c}} \delta[bin(\mathcal{Y}_D(\hat{s}, \mathbf{x}_{\mathrm{q}})) - \mathrm{i}] \delta[bin(\mathcal{Y}_D(\hat{s}, \mathbf{x}_{\mathrm{r}})) - \mathrm{j}] \delta[bin(\mathcal{Y}_D(\hat{s}, \mathbf{x}_{\mathrm{u}})) - k] \\ & \delta\left[bin\left(tan^{-1}\left(\frac{y_{\mathrm{q}} - y_r}{x_{\mathrm{q}} - x_{\mathrm{r}}}\right)\right) - \theta\right] \delta\left[bin\left(tan^{-1}\left(\frac{y_{\mathrm{r}} - y_{\mathrm{u}}}{x_{\mathrm{r}} - x_{\mathrm{u}}}\right)\right) - \vartheta\right] \end{aligned} \tag{4.6}$$

where, for simplicity, the summations over the possible combinations of q, r, u, θ, and ϑ and have been replaced with c which exists in the 5-dimensional space of the elements of co-occurrence histogram b and θ and ϑ are the bin indices for the angle of orientation between $\mathbf{x}_{\mathrm{q}}$ and $\mathbf{x}_{\mathrm{r}}$, and between $\mathbf{x}_{\mathrm{r}}$ and $\mathbf{x}_{\mathrm{q}}$ respectively. q, r, and u, still adhere to the spatio-temporal constraints defined previously, and also illustrated in Figure 4.1(b). The bin number for the orientations are defined so that if the spatial distance between $\mathbf{x}_{\mathrm{q}}$ and $\mathbf{x}_{\mathrm{r}}$ is 0, the bin number

is also 0.

$$bin\left(tan^{-1}\left(\frac{y_{\mathrm{q}}-y_{\mathrm{r}}}{x_{\mathrm{q}}-x_{\mathrm{r}}}\right)\right)=\begin{cases}\frac{8}{\pi}tan^{-1}\left(\frac{y_{\mathrm{q}}-y_{\mathrm{r}}}{x_{\mathrm{q}}-x_{\mathrm{r}}}\right) & \text{if}|\mathbf{x}_{\mathrm{q}}-\mathbf{x}_{\mathrm{r}}|\geq 0\\ 0 & \text{otherwise}\end{cases} \tag{4.7}$$

and $bin\left(tan^{-1}\left(\frac{y_r-y_{\mathrm{u}}}{x_r-x_{\mathrm{u}}}\right)\right)$ is similarly defined.

Let us call the set of spatio-temporal positions and saliency values which correspond to each element of the 5-dimensional co-occurrence histogram b, a triple. That is, each co-occurrence in b is accompanied by a representation of the likelihood of 3 saliency values co-occurring within a local neighbourhood, defined by the spatio-temporal relations in Figure 4.1(b). The triple feature vector, $\mathbf{f}_{\mathcal{Z}}=[\mathbf{f}_{\mathrm{q}}\ \mathbf{f}_{\mathrm{r}}\ \mathbf{f}_{\mathrm{u}}\ t]^{\mathsf{T}}$ contains 3 sets of attributes corresponding to 3 spatio-temporally connected locations in 3 consecutive frames the time frame at $\mathbf{x}_{\mathrm{q}}$ such that $\mathbf{f}_{\mathrm{q}}=[\mathbf{x}_{\mathrm{q}}\ \hat{s}\ \mathcal{Y}_D(\hat{s}\ \mathbf{x}_{\mathrm{q}})]^{\mathsf{T}}$ where $\mathbf{x}_{\mathrm{q}}=[x_{\mathrm{q}}\ y_{\mathrm{q}}\ t]^{\mathsf{T}}$, and for less cluttered notation the salient spatio-temporal peak $\hat{s}=[\hat{s}_{\mathrm{s}}\ \hat{s}_{\mathrm{t}}]^{\mathsf{T}}$ [1]. The term $\mathcal{Y}_D(\hat{s},\mathbf{x}_{\mathrm{q}})$ is again the temporal saliency at the spatial location $\mathbf{x}_{\mathrm{q}}$ and scale, $\hat{s}$ in the current frame. $\mathbf{f}_{\mathrm{r}}$ and $\mathbf{f}_{\mathrm{u}}$ are similarly defined but $\mathbf{x}_{\mathrm{r}}$ is either the same or one of the neighbouring 8 pixels of $\mathbf{x}_{\mathrm{q}}$ in the previous frame and $\mathbf{x}_{\mathrm{u}}$ is similarly defined in relation to $\mathbf{x}_{\mathrm{r}}$. Therefore, for element b_{c}, a set of corresponding feature vectors exist $\mathbf{F}_{\mathrm{c}}$ which contains all $\mathbf{f}_{\mathcal{Z}}$ for this element.

4.2 Temporal binding

The previous section described a representation of the likely local motion of a single object. However, higher semantic levels of inference are needed in order to understand interactive behaviour or cause-effect behaviour amongst multiple agents. While the common assumption is that interactions can only occur at close proximity, (Galata et al., 2002; Hogg et al., 1998; Johnson et al., 1998; Oliver et al., 2000; Park and Aggarwal, 2003) in reality, it is possible to observe interactions from highly spatially separated objects (Gong and Xiang, 2003). In order to detect interactions that occur at any spatial distance, it is necessary to represent temporally correlated but spatially separated behaviour. Temporal correlation can be achieved by finding the simultaneous co-occurrence of multiple changes of salient motion

[1] They will be written with the relevant spatial or temporal subscript in cases where there may be ambiguity

where simultaneity is better defined more loosely as some local temporal neighbourhood. Instead of just accumulating co-occurrences of temporally salient regions, we can now use the more descriptive triples, defined in the previous section. Then, if these temporally correlated motion patterns are suitably unusual, we can define these as events.

By accumulating changes in the likelihood of a particular spatio-temporal configuration of saliency values over time, we can build models of the likely temporally co-occurring changes in the scene dynamics and therefore identify unusual spatially separated co-occurrences. To accumulate a model of temporally correlated behaviour, a 2-dimensional co-occurrence matrix or histogram n is created. n is formed by accumulating the co-occurring changes in $b(t)$ over a local temporal neighbourhood $\{\tau : t-2 \leq \tau \geq t\}$, which is weighted by a spatial overlap function β_{cg}.

$$n_{\mathrm{lm}}(t) = \sum_{\mathrm{c}} \sum_{\mathrm{g}} \delta[bin(b'_{\mathrm{c}}(t)) - \mathrm{l}]\delta[bin(b'_{\mathrm{g}}(\tau)) - \mathrm{m}]\beta_{\mathrm{cg}}, \mathrm{c} \neq \mathrm{g} \tag{4.8}$$

where c exist in the set of all elements in the 5-dimensional version of $b'(t)$ from Equation 4.6 , g exists in the set of elements in the 5-dimensional version of $b'(\tau)$, l and m exist in one of M equally distributed intervals, $b'(t)$ is simply a discrete differentiation of $b(t)$ over time such that $b'(t) = b(t) - b(t-1)$. A spatial overlap function β_{cg} is defined such that it minimises spatial overlap and therefore co-occurrences in n between the spatio-temporal locations of the triples in b. Spatial overlap is determined (between a pair of triples in 2 elements of matrix $b'(t)$) by whether their mutual spatio-temporal locations yield the same salient spatial scale and position. Triples are considered to be overlapping if 2 or more of their spatial coordinates and corresponding spatial scales are equal. So for 2 triples with corresponding sets of coordinates d and e respectively,

$$\beta_{\mathrm{cg}} = \begin{cases} 1 & \text{if } \ Overlap_{\mathrm{cg}} \leq 2 \\ 0 & \text{otherwise} \end{cases} \tag{4.9}$$

$$Overlap_{\mathrm{cg}} = \sum_{\mathrm{d}} \sum_{\mathrm{e}} \begin{cases} 1 & \text{if } ([x_{\mathrm{d}}\ y_{\mathrm{d}}]^{\mathsf{T}} = [x_{\mathrm{e}}\ y_{\mathrm{e}}]^{\mathsf{T}}) \wedge (\hat{s}_{\mathrm{s}}(x_{\mathrm{d}}, y_{\mathrm{d}}, t_{\mathrm{d}}) = \hat{s}_s(x_{\mathrm{e}}, y_{\mathrm{e}}, t_{\mathrm{e}})) \\ 0 & \text{otherwise} \end{cases} \tag{4.10}$$

where d indexes all the pixel locations in the current triple and e indexes all the locations in the corresponding co-occurring triple. Both $bin(b'_{\mathrm{c}}(t))$ and $bin(b'_{\mathrm{g}}(\tau))$ are defined similarly except now the maximum and minimum values define extremes in the changes of the co-occurrence histogram b over time, which are calculated from all possible values in $b'_{\mathrm{c}}(t)$ where t exists for all frames of the sequence that are considered.

$$bin(b'_{\mathrm{c}}(t)) = b'_{\mathrm{c}}(t) \frac{M}{max(b'_{\mathrm{c}}(t)) - min(b'_{\mathrm{c}}(t))}, \text{ where } bin(b'_{\mathrm{c}}(t)) \in \mathbb{I} \tag{4.11}$$

To reduce computational complexity, not all corresponding coordinates in $\mathbf{F}_{\mathrm{c}}$ are checked for overlap. It is assumed that a selection is representative of the whole group. We are able to make this assumption since the idea is that every element in b can be treated as a group of similar activities. For example, in a busy motorway, it doesn't matter if a car is in the left or right lane since its descriptor will be the same as long as both lanes of the road go in the same direction. Ideally, we would like to treat all activity in these lanes as the same, while traffic coming from the other direction should be treated differently. Allowing this level of spatial ambiguity means that populating a model of a particular class of motion is more readily achievable with fewer frames.

For each element of $b'(t)$, a set of feature vectors exist, which represent the pairings of different triples. So for the feature vector of a triple pair, $\mathbf{ff}_{\mathcal{Z}\mathrm{lm}} = [\mathbf{f}_{\mathcal{Z}\mathrm{c}}\ \mathbf{f}_{\mathcal{Z}\mathrm{g}}]^{\mathsf{T}}$ and the set of all $\mathbf{ff}_{\mathcal{Z}\mathrm{lm}}$ is stored in the triple pair feature vector $\mathbf{FF}_{lm}$.

4.3 Quantifying salient correlations

Now that we have a method of describing local spatio-temporal configurations of saliency values and how they co-occur, the next step is to measure how salient co-occurrences are in

order to find unusual ones. We can do this by finding unusual fluctuations in $n(t)$, defined in Equation 4.8 since pairs of triples which are less well modeled are likely to be caused by unusual or salient behaviour.

Unusual co-occurrences are found by accumulating the frequency of occurrence of $n'(t)$ where $n'_{\mathrm{lm}}(t) = n_{\mathrm{lm}}(t) - n_{\mathrm{lm}}(t-1)$ over the whole sequence into a 1-dimensional histogram o. Therefore, the change in $n(t)$ is used to define whether the set of co-occurring triples are considered interesting or not. If a particular level of change occurs frequently, we can view this as any type of non-salient activity. However, if it occurs less often, then it is more likely to be salient. An element of the co-occurrence histogram o is defined as

$$o_{\mathrm{v}} = \sum_{t} \sum_{\mathrm{f}} \delta[bin_{\mathrm{f}}(n'(t)) - \mathrm{v}] \tag{4.12}$$

where v exists in 1 of N bins, f indexes all the elements in $n'(t)$ and t exists in the set of all frames in the sequences that is being considered. Here, we accumulate the frequency of certain levels of change in $n(t)$ over time. $bin(n'(t))$ is defined in a similar way to Equation 4.11 except now the maximum and minimum is taken from all values of $n'(t)$ over the whole image sequence that is considered.

4.4 Algorithm Summary

A diagrammatic representation of the co-occurrence algorithm is shown in Figure 4.2. Part (a) shows the initial feature extraction stage described in the previous chapter except now the spatio-temporal kernel has been modified to be one-sided to reduce temporal ambiguity. At each frame, the temporally salient responses for each peak in entropy are concatenated into a feature vector $\mathbf{f} = [\mathbf{x}\ \hat{s}\ \mathcal{Y}_D(\hat{s}\ \mathbf{x})]^{\mathsf{T}}$. Then , in part (b), the co-occurrence histogram b is formed, as described in Equation 4.5 but with the addition of 2 extra dimensions to account for the relative spatial orientations of the temporal saliency values, as shown in Figure 4.1(c). As mentioned before, each element of b has an associated feature vector $\mathbf{f}_{\,3}$ corresponding to the 3 feature vectors of its triple. By accumulating b over time, we have a snapshot of

b at each time frame which can be used to monitor the evolution of its elements over time and how they co-occur with other triples. Therefore, in part (c), the change in b over time is accumulated in a co-occurrence histogram n, as described in Equations 4.8-4.10. In part (d), the one-dimensional co-occurrence histogram o is created using Equation 4.12. In order to find the most salient co-occurrences from o, its bins must be reordered in ascending frequency. After this, a small percentage of the least frequent co-occurrences are selected from the ranked values in o.

4.5 Temporal ambiguity of the temporal saliency algorithm

The feature extraction method outlined in the previous section was modified to a one-sided version as shown in Figure 4.3 to remove the ambiguity from the selected peak in temporal scale. Equation 3.11 is modified so that every increase in temporal scale s_t leads to widening of the spatio-temporal cylindrical kernel in one direction of time. In this case, all the experiments carried out here assume that the kernel is expanded backwards in time. So the shaded area in Figure 4.3 represents the kernel at the lower temporal scale while the entire volume represents the kernel at the next scale up.

4.6 Interactive phenomenon of a bouncing ball

In order to demonstrate the significance of what the overall algorithm can do, let us observe a simple experiment of a sequence of a bouncing ball. Just under one full cycle is used for this experiment as shown in the leftmost column of Figure 4.4.

The experiments in this section are split into 2 different methods of finding salient regions of motion. The first uses the spatio-temporal binding method described in Equation 4.3 where co-occurrences of saliency values are accumulated between pixels in a local spatio-temporal neighbourhood only. To view the most salient co-occurrences from this method, the co-occurrence histogram b from Equation 4.3 is concatenated into a single 1-dimensional histogram and then reordered in ascending frequency. A percentage of the bins with lowest frequency are labeled as the most salient local spatio-temporal co-occurrences. Let us call

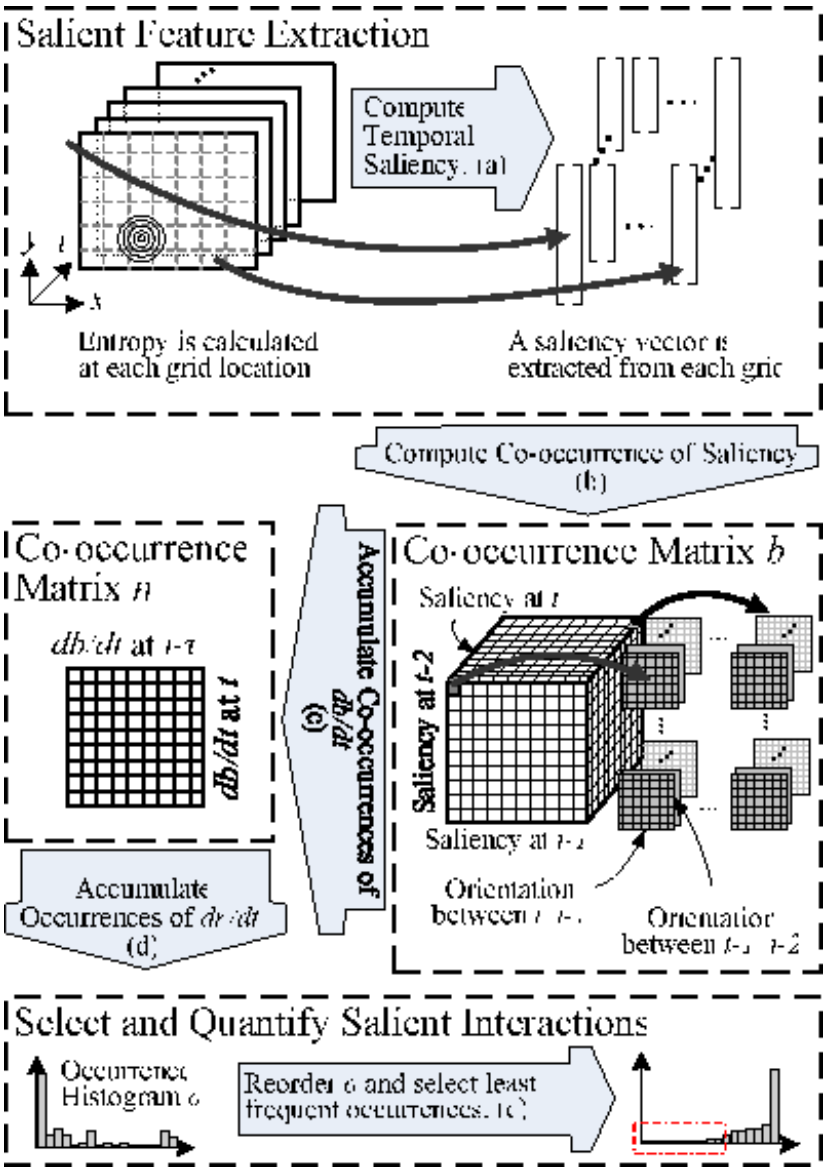


Figure 4.2: Diagrammatic representation of the algorithm.

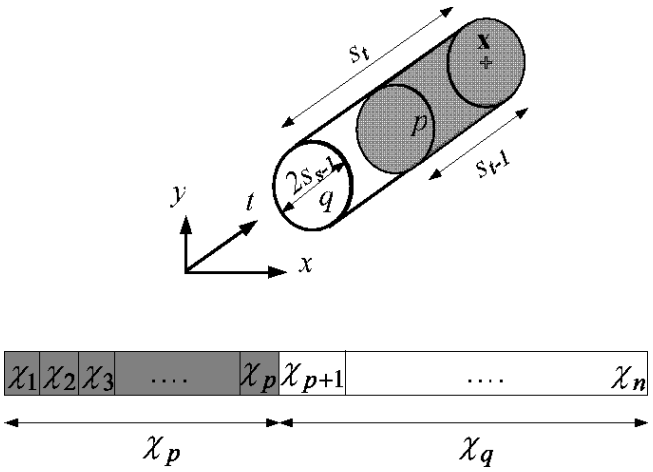


Figure 4.3: One sided sampling kernel for calculating entropy variation over spatio-temporal scales.

this the basic method.

In the second set of experiments, the full algorithm described in Section 4.4 is implemented and shown to have much more compelling results than using local spatio-temporal binding alone. Here, the results are visualised by firstly selecting the most salient bins from the histogram o, defined in Equation 4.12. Using the feature vector $\mathbf{FF}_{\text{cg}}$ associated with each of these salient bins. Let us call this the complex method.

In row (a) of Figure 4.4, some example frames of a bouncing ball sequence are shown. The actual length of sequence that was used for this simple experiment involved just under one cycle of the ball moving up and then down again. The frames shown in Figure 4.4 show the entire sequence but sub-sampled by 1. The frames in row (b) show the salient frames, temporally aligned to the original sequence in row (a), that were highlighted using the basic co-occurrence method described in Section 4.1. The top 10% most temporally salient regions are shown and clearly clustered around the turning point of the motion of the ball. The circles indicate temporally salient regions of the frames where the size of the circle indicates the spatial scale at which the entropy peaked in space and time and its brightness is proportional to the saliency of the region.

In the next row, the complex method, as described in Section 4.4 was applied to the original sequence in order to find salient spatially separated but temporally correlated motion. In this

Figure 4.4: A bouncing ball phenomenon : A simple experiment showing the comparison of using the basic co-occurrence method for correlating local temporally salient regions and also the higher level spatially separated co-occurrence of salient regions. Row (a) shows the original sequence, which has been sub-sampled by 1. Row (b) shows the salient frames that were found using the basic co-occurrence method, as described in Equation 4.3 which are temporally aligned with the frames in row (a). The white circles represent temporally salient spatial locations accumulated over time. The size of the circles indicate the spatial scale at which the spatio-temporal entropy-scale characteristic peaked and the brightness of the circle indicates higher temporally saliency. Row (c) shows the most salient frames taken using the method described in Section 4.4 where temporal alignment with the original sequence in row (a) is again utilised. Here, the black lines connect together temporally salient, spatio-temporally neighbouring locations or a triple. The grey lines represent the correlation between spatially separated triples. The thickness of the line shows how salient the interaction between the regions are.

case, the circles are as described for the previous set of results, black lines indicate which neighbouring regions are temporally salient triples and the grey lines show the correlation between spatially separated but temporally correlated sets of triples where the thickness indicates how unusual the interaction is. Here, the results indicate an unusual correlation between the ball and its shadow when the ball and shadow both change direction. This shows quite powerfully, the purpose of finding spatially separated correlated motion since the ball and shadow are inextricably linked whilst being significantly spatially separated in the frame.

On comparison of the results using the basic and complex method, we can see that using local spatio-temporal binding alone does not necessarily give us the full picture of the temporal dynamics within a scene. At this stage, it could be argued that the local spatio-temporal binding process simply measures unusual directions of motion and perhaps it would seem that using other more straightforward motion direction estimators would be more appropriate. However, the advantage of accumulating co-occurrences of local spatio-temporal saliency values is that this descriptor is not limited explicitly to estimating motion direction. If the scene is more cluttered, it could be that neighbouring responses exhibit different characteristics which we might consider more as a local spatio-temporal texture. Under cluttered conditions, estimating direction of motion is more difficult due to multiple occlusions but even if a scene is cluttered, this should not make the motion patterns within it indistinguishable. Accumulating co-occurrences of the temporal saliency from local spatio-temporally connected regions means that a model of different configurations of temporal saliency can be extracted to represent some temporal dynamics within the scene without performing any object tracking.

The conclusions drawn from this experiment are that the basic method is only able to detect salient local changes in motion for single objects while the complex method can find relations between them. This sort of behaviour is particularly useful in natural scene data.

4.7 Experiment

Experiments were carried out on 2 contrasting scenes: a busy traffic scene and a corridor entrance scene. Typical frames of the 2 scenes are shown in Figure 4.5.

4.7.1 Basic co-occurrence

The co-occurrence of saliency features using Equation 4.3 and illustrated in Figure 4.1(a) was calculated for a busy traffic scene containing 3100 frames captured at 25Hz, and sub-sampled by 5 frames where the accumulated results were spread between 10 bins. Most of the single co-occurrences were attributed to the spatio-temporal location of the second reversing vehicle in the sequence as shown in Figure 4.6 where a selection of the frames highlighting the 10% least frequent co-occurrences of matrix $b(t)$ are shown where $b(t)$ was calculated using Equation 4.3. Figure 4.6(a) shows a car slowing down and changing lanes. The rest of the figure shows a detected instance of the second reversing car incident where the car in question has been manually highlighted by a white box for clarity. Here we can see that the method can detect when the car has been stationary for an unusually long period of time before it starts to reverse.

In the sequence that was considered, there were 2 different reversing car scenarios and results have highlighted the second chronological one. There were difficulties detecting the first reversing car in the sequence since the relative motion was less because it was further away from the camera. Key frames showing this reversing car are shown in Figure 4.7.

Figure 4.8 shows that despite a fairly uniform sampling of the scene, the area closer to the camera yields, on average, much higher saliency values simply due to the effect of perspective. That is, objects closer to the camera exhibited higher between pixel variations over time and hence higher temporal saliency. Therefore, the spatial location of the first (undetected) car reversing event suffers from low saliency values. Although our method did detect and register a high saliency value at the location where the vehicle, shown in Figure 4.7, stopped and started reversing (see top left corner in Figure 4.8), this was not considered by the algorithm to be globally salient. This suggests that performing approximate local spatio-

Figure 4.5: Typical frames from the 2 scenes. Top 2 rows : busy traffic. Bottom2 rows: corridor entrance.

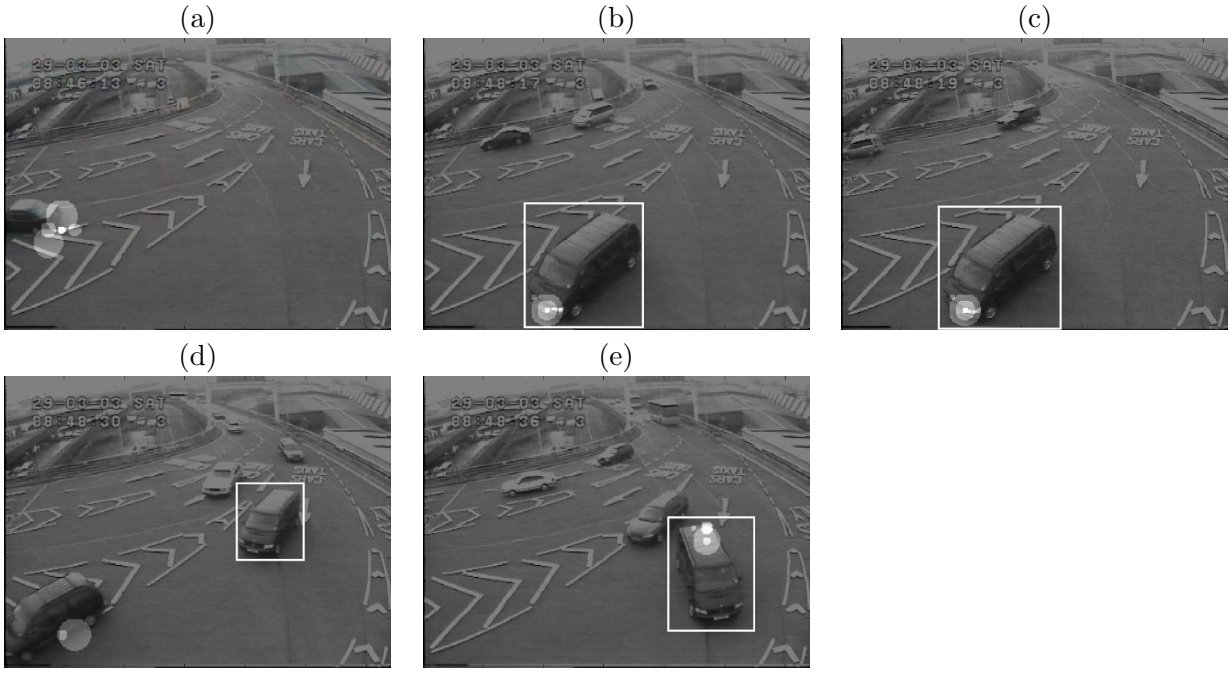


Figure 4.6: Frames illustrating the ranked top 10% least frequent co-occurrence elements from a busy traffic scene using the basic method from Eqn. (4.3) where the reversing car is highlighted manually by a white box. Co-occurrences at particular spatial locations are highlighted with circles. Sizes of the circles indicate the scale at which the entropy peaks within that local spatio-temporal neighbourhood and their intensity represents the temporal saliency where lighter shades represents higher values.

temporal binding is not enough to distinguish between salient and non-salient activities. However, we aim to detect all types of salient events with equal saliency values regardless of perspective scale. It follows that accumulating co-occurrences between salient spatially separated spatio-temporal triples would remove sensitivity to perspective variations.

It is important to note that representing the local spatio-temporal configuration of saliency values will only help to discriminate between different types of salient motion texture and is not the same as merely measuring changes in the motion direction. We refer to this configuration structure as a texture since there is no guarantee that the co-occurrence of local spatio-temporal saliency values will correspond to exclusive statistically usual directions of motion. The co-occurrence histogram can only capture this information if and only if there is relatively little background clutter such as in the bouncing ball example in Figure 4.4. In all other cases, the configuration of local spatio-temporal saliency values can only be considered

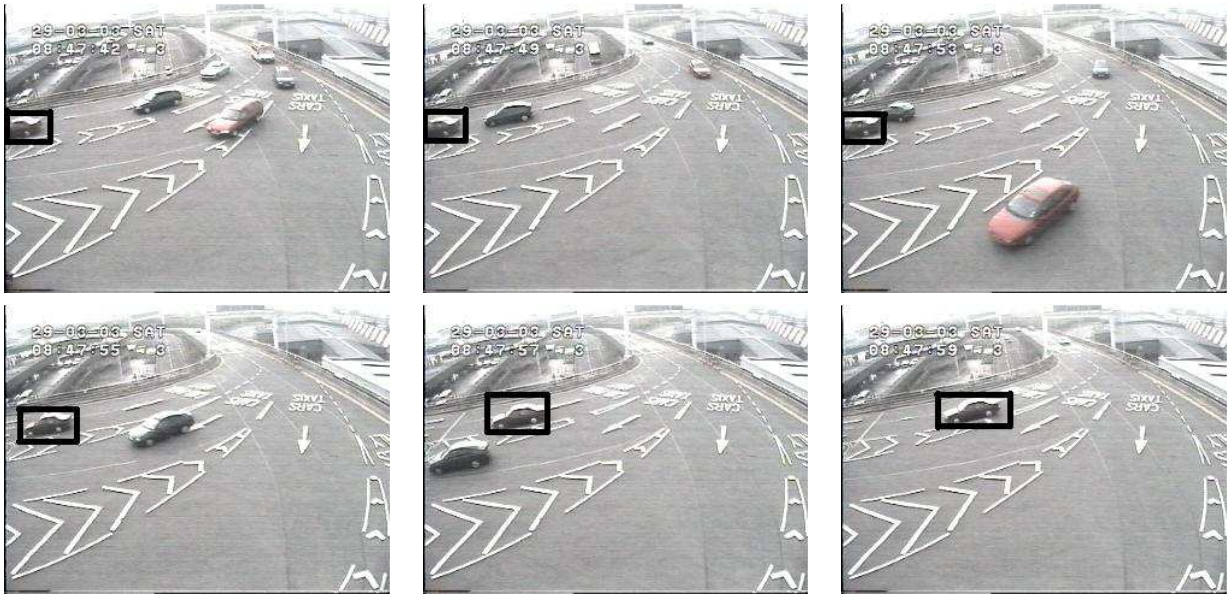


Figure 4.7: Key frames of the first of 2 different reversing cars scenarios in the busy traffic scene. At first the car stops for a while and causes another to stop and change lane. Then the car reverses. The car has been manually labeled with a black box in each frame.

as a motion texture.

Similar results are shown for a scene of a secure entrance in a corridor, containing quite complex motion patterns to 3 possible entrances/exits. A sequence consisting of 4000 frames taken at 10Hz, and sub-sampled by 3 frames was used for our experiment. Figure 4.9 shows the spatio-temporal location of the least frequent co-occurrences of saliency where row (b) shows the false positives and the row (c) shows the true positives for this sequence. Many false-positives were caused by changes in intensity when the doors were opened and closed. People were also highlighted since the path they took within the scene, their height, or intensity of clothing were unusual. Whilst this might indicate that more data is needed, we show later that these anomalies are removed with higher levels of co-occurrence. The true-positive results, in Figure 4.9(b) shows a person running to catch the door, and also 2 people who were unable to go through and turned back.

Our experiments on outdoor and indoor scenes show good results for detecting salient (unexpected) motion patterns (e.g. reversing car or people turning around / wandering in front of an entrance because they cannot open a door). However, using the basic method,

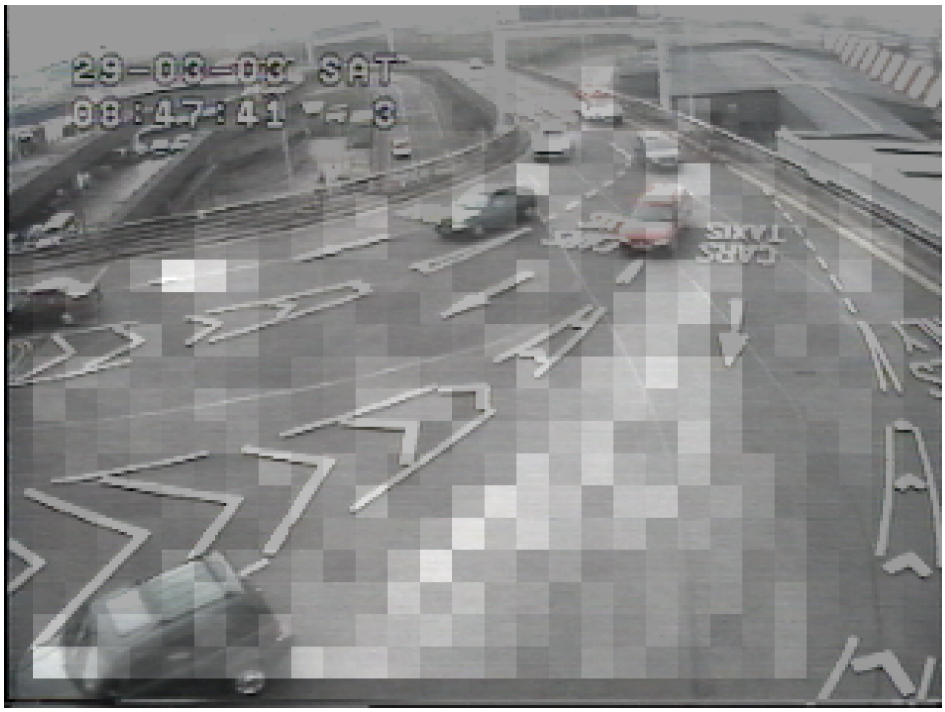


Figure 4.8: Mean saliency values over time for the traffic scene. Lighter areas show higher saliency. This highlights the problem of biased saliency values due to perspective where cars further in the distance will have a lower temporal saliency just because their relative motion is smaller. However, using the complex method, this problem is reduced since co-occurrences between spatially separated regions are insensitive to how salient a single spatio-temporal region is. Considering co-occurrences of otherwise less temporally salient motion means that the complex method quantifies the saliency of the interaction itself.

we can only detect unusual changes to local spatio-temporally connected salient regions. However, there are many cases where the interaction between different objects is more salient in combination. If we could model these interactions in terms of temporally correlated changes in the motion patterns, then it would be possible to identify scene dynamics that have a more complex semantic meaning.

4.7.2 Temporal correlation and quantification

The method, as described in Section 4.4 and Figure 4.2 was employed, using 20 bins to accumulate correlations of the rate of change of co-occurrence of saliency values b'. Rather than assuming temporal correlations occurred simultaneously, a temporal interval of 3 sub-

Figure 4.9: Frames illustrating the ranked top 10% least frequent co-occurrence elements from a corridor and entrance scene using the basic method from Eqn. (4.3). Co-occurrences at particular spatial locations are highlighted with circles. Sizes of the circles indicate the scale at which the entropy peaks within that local spatio-temporal neighbourhood and their intensity represents the temporal saliency where lighter shades represents higher values.

sampled consecutive frames was used. For complexity reasons, the algorithm was streamlined by not taking into account all the normal fluctuations to matrix n. The ratio of usual-unusual fluctuations were high enough that the results would not be affected by ignoring some normal data. The top graphs in Figures 4.10 and 4.11 shows the maximum of matrix n' at each frame for the busy traffic scene and corridor scene respectively, as described in Section 4.2. Peaks show salient spatially separated but temporally correlated motion.

For the busy traffic scene, most of the occurrences were grouped into one bin and the other frequency values were at least 2.5×10^{-05} times smaller. Only 9 out of 20 bins were filled. The smaller values are shown in Figure 4.10. The second reversing car event is detected very clearly in Figure 4.10(h-j) where salient correlations were found between the reversing car (that was slowing down before reversing), marked by a blue box, and other normally behaving cars within the scene. The first (previously undetected) reversing car marked by a red box, is still not detected but a car that has to slow down and change lanes due to the reversing car, marked by a green box, is detected. Salient correlations were found between the affected car and surrounding cars that were behaving normally. This is illustrated in Figure 4.10(d-f) where the car marked with a green box, is slowing down and has been correlated with area in the background, which corresponds to a faster moving car. Some of the peaks in the top graph of Figure 4.10 correspond to the 2 reversing car events have been marked.

Figure 4.10(a-b) highlights the event of the middle car, marked with a green box, changing lanes. However it has to slow down due to the car on the left. The car on the left also slows down and is therefore correlated with a normally behaving lorry moving down the frame in a road at the very top. The location of the lorry is marked by a cyan box at the top of Figure 4.10(a,b) which has been obscured by the visualisation of the results. Unfortunately, the car on the left marked with a red box, is also correlated with itself 3 frames previously. Figure 4.10(j) also highlighted problems caused by the assumption that no triples should overlap if the corresponding spatial scales at which the entropy peaked were the same since some spatial overlap was possible. However, in this case, the response was caused by the motion of the same car correlated with itself a few frames earlier. To overcome this, more specific object detection would be required. The thickness of the lines between interacting

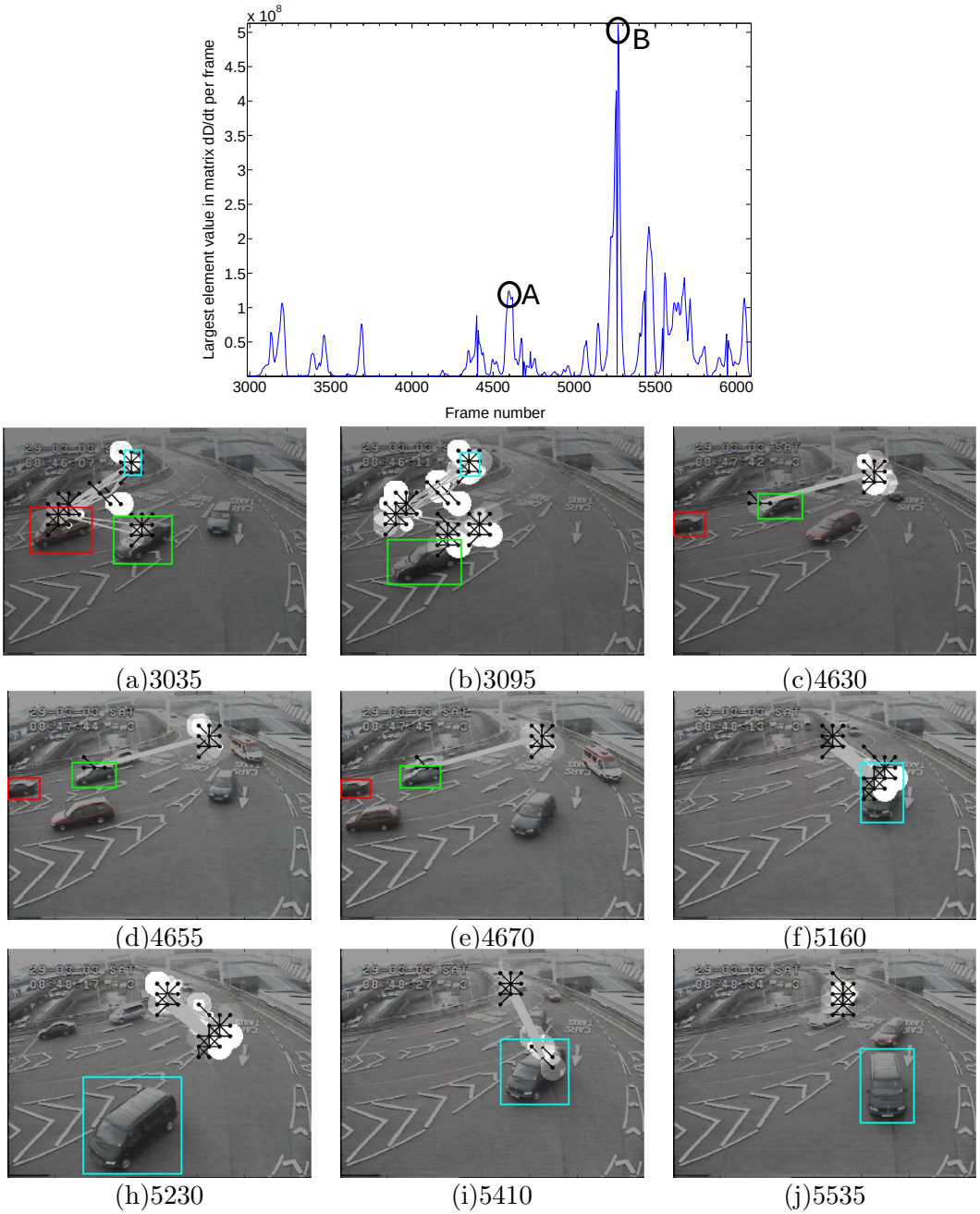


Figure 4.10: Results from applying the complex method to the busy traffic scene. The top graph shows salient temporally correlated motion events in each frame. The highlighted peaks show the frames containing the 2 reversing car events. Peak A:(undetected) reversing car event. Peak B: detected reversing car event. (a-j) shows frames of the corresponding triples of the least frequent elements of histogram o. Salient co-occurrences are highlighted with circles. Sizes of the circles indicate the scale at which the entropy peaks within that local spatio-temporal neighbourhood and their intensity represents the temporal saliency where white represents the highest value. Black lines show the salient triple within a local spatio-temporal neighbourhood. Grey lines show correlations between spatially separated salient motion where thickness is proportional to saliency of the interaction. The corresponding frame numbers are shown under (a-j).

objects provides a clearly distinguishable difference between more and less salient interactions.

The results from applying the complex method to the corridor scene are shown in Figure 4.11. The scene contained periods of no activity, where low temporally salient responses were still detected due to the noise in the image frames. This meant that all events involving opening the secure doors were detected as salient, as shown by the large numbers of peaks interspersed by periods of no activity in the top graph of Figure 4.11.

The 2 salient motion events involving a person running to catch the door and also 2 people who are unsuccessful in entering, which were detected in the previous section were also detected using this method, as shown in Figure 4.11(q-s), which were amongst the least frequent occurrences. The first salient event where someone rushed to catch the door, is shown in (o,p) where correlations were confined to the area at the top of the frame, rather than between running person and the closing door. However, now a salient interaction detected between the door and the parts of the scene beyond the gap in the door, as shown by the grey lines. The second correctly detected salient event was the 2 people who cannot get in and turn round in (q-s). The correlated activity between the heads of the 2 people was detected, as shown in (r,s). Note that the interaction between the heads was considered less salient than loitering round the door for an unusual amount of time since the grey lines in (q,r) are thicker than those in (s).

Many correlations were made between the person opening the door and the motion of the door, such as those shown in (k,l,n). Again, there were some correlations of the motion of a person with themselves in later frames in all the examples of Figure 4.11. In (k,m), correlations were also made between the person and their reflection in the glass of the secure doors.

Overall, the results using the complex method show that discrimination between salient activities was apparent, though better discrimination between different interactions would be needed to facilitate a more complex and informative hierarchical structure. It is important to note that all the explanations of the results provided in this section are only provided through a manual interpretation of the responses given the activity in the scene. For a fully automated system that was able to construct more specific semantic meaning from the

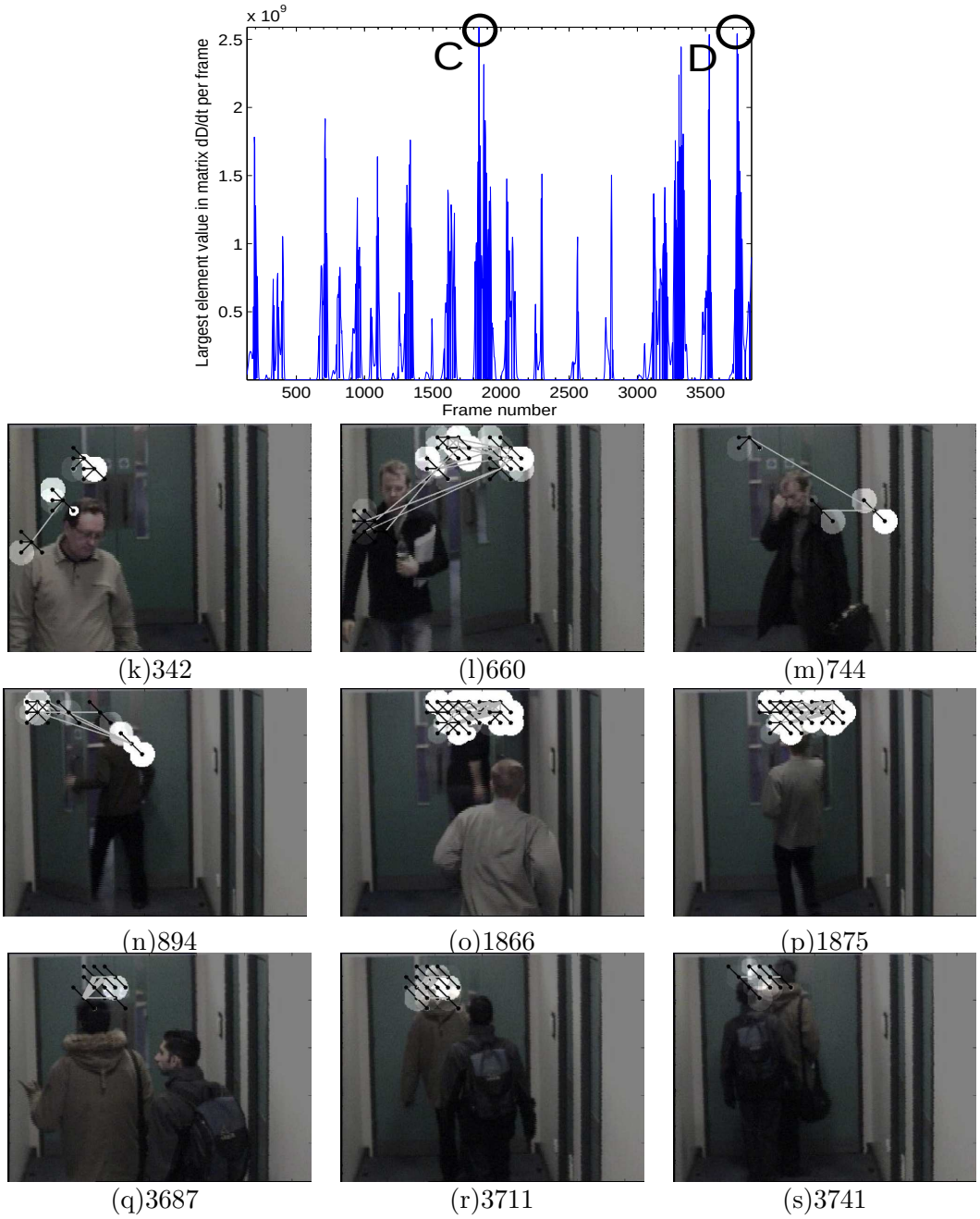


Figure 4.11: Results from applying the complex method to the corridor scene. The top graph shows salient temporally correlated motion events in each frame. The peaks which are highlighted, show the frames containing the 2 reversing car events. Peak C: person running to catch the door. Peak D: 2 people who cannot open the door and turn back. The rest of the figure shows frames of the corresponding triples of the elements of histogram o with least frequent elements. Co-occurrences at particular spatial locations are highlighted with circles. The sizes of the circles indicate the scale at which the entropy peaks within that local spatio-temporal neighbourhood and their intensity represents the temporal saliency where lighter intensity represents higher values. Black lines show the each triple generated within a local spatio-temporal neighbourhood. Grey lines show correlations between spatially separated salient motion where thickness is proportional to saliency of the interaction. The corresponding frame number is indicated under each frame.

activities, it would be necessary, at the first stage, to apply object detection to the regions that are highlighted. In order to describe the interactions, top-down knowledge would need to be applied. It is worth noting at this stage, that all detections that have been reported in this section were described based on the spatial overlap of the co-occurred regions and the particular objects of interest. In order to automatically generate semantic meaning from the responses would require the objects to be recognised and the individual activities to be classified, which has not been covered in this chapter. However, this will be approached in more detail in the next chapter.

4.8 Discussion and conclusions

In this chapter, the power of using contextual information for representing and extracting low and mid level features has been demonstrated. Using the measure of saliency described in Chapter 3, it has been possible to find relations between spatially local as well as separated regions using just their temporal saliency measure. The results demonstrate that being able to find temporally correlated relations between spatially separated regions provides a powerful method of modeling the dynamics within a scene. More importantly, the results show that significant high-level dynamics of a scene can be detected from the scene data alone. Earlier in the chapter, I argued that there was no need to select model order for co-occurring data. Whilst choosing the number of bins may be considered as manual model order selection, the same number of bins was used for 2 contrasting scenarios where the depth of perspective, the size of the moving objects, and typical object trajectories were very different. Therefore as long as the bins are chosen to avoid sparsity issues, the method works fairly well.

A weakness of the method is that it is not able to detect unusual behaviour by an individual object. That is to say that the temporally salient responses do not necessarily highlight an entire spatially homogeneous object. Rather, the part which exhibits consistent, non-cyclic temporal change is highlighted in each case. So in most cases, only the leading edge of the moving object is detected. Applying the overlap constraint in Equation 4.10 could provide only very loose exclusivity between two triple responses that were partially spatially

connected. However, even in the very simple scenario of a bouncing ball, the results in Figure 4.4 show that many salient triples were identified found around different parts of the the moving ball. For more robust identification of objects, more specific constraints need to be applied to the temporal saliency algorithm to ensure that more salient regions are found. The problem with just detecting a peak in spatial saliency is that this can refer to the change in entropy as the kernel crosses the boundary between the ball and the background, which does not require the circular kernel to be centred at the centre of the salient object.

However, in most practical surveillance situations, salient activity is only significant or worth adding resources to, when the activity has a direct impact on other normal activity within the scene. Certainly, in surveillance scenarios, one could imagine that a single person behaving in an anti-social manner is not considered so much of a hazard unless they start to antagonise other members of the public.

Overall however, the implicit assumption of the algorithm described in this chapter is that the temporal dynamics are the most significant part of understanding collective behaviour. That is, relative spatial location becomes much less important than one might have envisaged, which proves that even significantly spatially separated activities can be influenced by each other. Although the results have been considering collective behaviour from humans, there are still more cues that have been ignored for better understanding of these types of complex behaviours. Within a single object, other pose-based cues, such as human gestures, may provide more means of understanding interactive behaviour from a more sophisticated event-based structure. Using such methods will reduce the heavy and almost exhaustive computation required for the algorithm described in this chapter.

5 Spatio-temporal saliency and temporally correlated human interaction

In this chapter, we return to classifying natural human gestures except now, they are classified within the context of human-human interaction. Chapters 3 and 4 referred to methods for extracting temporally salient features which could be correlated so that salient temporally correlated motion could be identified. We bring all these ideas together for this problem and make some further modifications to the temporal saliency algorithm.

In keeping with the previous chapter, the novelty of this approach is that the interacting people can be significantly spatially separated, which is contrary to research in this area, which tends to make the assumption that interactions are bound to occur within a tight local spatial neighbourhood (Galata et al., 2001, 2002; Park and Aggarwal, 2003; Hogg et al., 1998; Oliver et al., 2000). Since we can no longer rely on spatial proximity as the primary cue for interaction, temporal correlation of salient motion or pose information becomes much more significant. Indeed, finding a suitably rich and discriminative representation of the local motion of the human body as opposed to the macro motion, becomes much more significant. By using the terms 'local' motion, I refer to the motion of the limbs of the body whilst the person is standing or walking whilst 'macro' motion refers to the overall motion of the person (i.e. whether they are standing and therefore stationary, or walking).

In this chapter, a novel representation of shape is proposed. However, in order to evaluate the effectiveness of such an approach, an extension to the idea of finding temporally correlated behaviour is approached by using a more formulated method to identify likely co-occurrences of pose or configuration features between two interacting (or non-interacting) people.

5.1 Finding temporally salient spatially homogeneous regions

In the previous two chapters, the temporal saliency was calculated and then treated as a measure of the local spatio-temporal characteristics of a video sequence. However, this does not necessarily provide the best description of the salient region since it is likely that a single motion pattern will trigger many salient responses for a spatio-temporal kernel centred around the same spatial location but at a neighbourhood of different time frames. This reinforcement is an interesting and useful property that would be useful to exploit in order to perform more spatio-temporally specific foreground/background separation. Furthermore, the method in its current state, is unable to run in an on-line form and requires some modifications so that we can move closer to more practical systems. While this will remove the global context properties of the temporal saliency algorithm, it will allow some form of salient homogeneous object detection.

The one-sided temporal saliency algorithm, which was implemented in Chapter 4 was applied to a busy traffic scene. A saliency map $\mathbf{G}(t)$ of each frame was created by accumulating all the saliency values in temporally salient cylindrical regions which overlapped each pixel in space and time. A resultant frame is shown in Figure 5.1 but shows a distinct problem since all the temporally salient regions identify areas of motion ahead of the motion of the object. This is one of the challenges that will be considered in the following sections where we approach the problem of extracting salient human forms from a sequence.

5.2 Extracting salient human forms from a cluttered scene

I have deliberately chosen not to use prior models of humans to detect single or groups of them since we are interested in making this method applicable to highly cluttered scenes where self occlusion occurs frequently amongst groups of people. The assumption is that most motion is likely to come from either background noise or people. Therefore, the temporal saliency algorithm defined in the previous two chapters is modified again to provide an effective method of extracting temporally salient homogeneous regions from a scene.

Figure 5.1: An example frame of the busy traffic scene computed using the accumulated temporal saliency algorithm. Temporally salient circular regions were accumulated over all relevant spatio-temporal locations. However, the temporally salient regions do not fully spatially overlap the moving objects.

5.2.1 Wide area scenes and low resolution

Let us firstly consider the case where the human form is represented by relatively few pixels in a wide area scene, which is common for sequences captured by outdoor surveillance cameras. Such a scene is shown in Figure 5.2. Here, we can see that the scene is extremely cluttered, and the people are represented by relatively few pixels, despite the high number of pixels which are used to represent the entire scene. The reason for choosing this scenario is that extracting the human form at lower resolutions is a simpler task since it is more difficult to differentiate between different parts of the body so it is easier to treat all the parts of the human form as a single object. There are, however, still challenges at this level, which would be interesting to explore.

At the first stage, temporally salient regions must be extracted from the scene. The cylindrical kernel which is used to capture the temporal saliency of a spatio-temporal region is only required to increase temporally backwards in time, as described in Chapter 4. This helps to eliminate temporal ambiguity when a peak in entropy at a particular temporal scale is found. However, here temporal saliency is only calculated at seed points which are spatially

Figure 5.2: An example of a highly cluttered wide area scene. The size of the image is 640 by 480 pixels but each person is only represented by about 20×40 pixels in width and height respectively. The scene also contains much clutter and there is the possibility of occlusion of the people from some parts of the scene.

salient. These seed points are extracted from a temporally differenced frame where each pixel is subtracted from an adapted median intensity value for that pixel. The reason for performing temporal differencing on the image first is that this significantly reduces the computation time of the algorithm, which is quite slow given the number of pixels per frame. The frame could have been sub-sampled but the rate was limited by the number of pixels which represented a single person so not much speed could be gained from sub-sampling alone. Furthermore, using temporal frame differencing still captures much background noise, where the most common method of removing clusters of very few pixels is by a manually set threshold (Needham and Boyle, 2003; McKenna et al., 2000; Stauffer and Grimson, 2000). However, no such threshold is used in this experiment and instead, such anomalies are eliminated by measuring and ranking the temporal saliency values at these seed locations.

Rather than finding spatially salient circular seed regions from the temporally differenced image, an elliptical kernel was used to approximate the dimensions of a person using a modification of the scale saliency algorithm by Kadir et al. (2004). This modified spatial

saliency was applied to the temporal difference image. The spatial scale term used in Equation (3.2) was altered with a shape vector $\mathbf{s} = [\alpha\ \rho\ s_s]^\mathsf{T}$ which represents the ratio, ρ, scale s_s and orientation, α, of the kernel.

$$\mathcal{H}_D(\mathbf{s}, \mathbf{x}) = -\sum_{d \in D} b(d, \mathbf{s}, \mathbf{x}) \log_2 b(d, \mathbf{s}, \mathbf{x}) \tag{5.1}$$

where only the spatial scale is varied while ρ and α remain constant to maintain a comparable distribution of radial overlap between consecutive scales. Therefore, only s_s is changed between scales for the intensity distribution $b(d, \mathbf{s}, \mathbf{x})$. Here, the ratio and angle of orientation of the ellipse is manually selected and maintained for the whole algorithm since it was assumed that most people in the scene are likely to be standing and therefore vertically oriented. It was found that any ratio that could encompass the whole human form proportionately well was fine. The inter-scale saliency, $\mathcal{W}$ and saliency measure $\mathcal{Y}$ are similarly modified to take into account the new measurements of change.

$$\mathcal{W}_D(\mathbf{s}, \mathbf{x}) = \frac{s_s^2}{2s_s - \Delta s_s} \sum_{d \in D} |b(d, \mathbf{s}, \mathbf{x}) - b(d, \mathbf{s} - \Delta\mathbf{s}, \mathbf{x})| \tag{5.2}$$

where the normalisation term $\frac{s_s^2}{2s_s - \Delta s_s}$ uses an approximation of the area of the ellipse based on the mean of the length of its major and minor which can be defined in terms of the scale and ratio as $s_s/\sqrt{\rho}$ and $s_s\sqrt{\rho}$ respectively, as defined in Kadir (2002). The saliency is then defined as before in Equation (3.7), which has been repeated here for convenience.

$$\mathcal{Y}_D(\hat{s_s}, \mathbf{x}) = \mathcal{H}_D(\hat{s_s}, \mathbf{x})\ \mathcal{W}_{D_{peak}}(\hat{s_s}, \mathbf{x}) \tag{5.3}$$

where the scales at which the entropy peaks are stored in the spatial scale vector $\hat{\mathbf{s_s}}$. These scales are identified using Equation (3.5).

At this stage, many of the spatially salient elliptical regions still identify areas of background motion since it has only been applied to the temporally differenced image. Therefore, the temporal saliency algorithm is applied to the original image sequence to suppress motion from dynamic background. Although the salient seed location is of an elliptical shape, it is

necessary to approximate the salient elliptical region using a circular kernel in order to prevent bias towards motion in the direction of the minor axis. This is particularly important for cases in the scene when someone is walking towards the camera and perspective distortion causes the vertical displacement to be smaller in proportion to the horizontal displacement of someone walking approximately parallel to the image plane.

The temporal saliency equation in this case, is slightly modified since it is not necessary to estimate the spatial scale as this has already been done. Therefore the temporal saliency is measured for a fixed spatial scale where the scale is large enough to encompass the whole of the salient ellipse. However, even if we select spatially salient regions from the temporally difference frame, the temporal saliency measure does not encompass how spatially salient the object is. It is quite possible that the people may remain relatively stationary for long periods of time and so we wish to relax the algorithm's sensitivity to salient motion alone and increase its sensitivity to spatially homogeneous elliptical forms. We could increase the maximum temporal scale of the kernel in order to cater for objects that remain stationary for long periods of time. However, this will severely increase computational complexity for only better performance in a few cases. Therefore, relaxing the sensitivity to salient motion alone seems sensible.

Firstly, the temporal saliency measure is defined to take into account the temporal scale at which the entropy peaks, $\hat{s_t}$ at the already selected spatial scale $\hat{s_s}$.

$$\hat{s} = \left\{ [\hat{s_s} \ \hat{s_t}]^{\mathsf{T}} : (\mathcal{H}_D([\hat{s_s}, \hat{s_t} - 1], \mathbf{x}) < \mathcal{H}_D([\hat{s_s}, \hat{s_t}], \mathbf{x})) \wedge (\mathcal{H}_D([\hat{s_s}, \hat{s_t}], \mathbf{x}) > \mathcal{H}_D([\hat{s_s}, \hat{s_t} + 1], \mathbf{x})) \right\} \tag{5.4}$$

Again, since it is likely that there will be more than one peak in entropy over spatio-temporal scale-space at a particular point $\mathbf{x} = [x \ y \ t]^{\mathsf{T}}$ in space-time, then a set of all spatio-temporal scales, $\hat{s} = [\hat{s}_s \ \hat{s}_t]^{\mathsf{T}}$, at which the entropy peaks is created which is denoted $\hat{\mathbf{s}}$. The spatio-temporal saliency is then defined using a weighted sum of the temporal saliency and the spatial saliency where the temporal saliency is defined

$$\mathcal{Y}_D(\hat{s}, \mathbf{x})_t = \mathcal{H}_D(\hat{s}, \mathbf{x}) \ \mathcal{W}_D([\hat{s}_s \ \hat{s}_t]^{\mathsf{T}}, \mathbf{x}) \ \mathcal{W}_D([\hat{s}_s \ \ \hat{s}_t + \Delta s_t]^{\mathsf{T}}, \mathbf{x}) \tag{5.5}$$

where both inter-scale functions have a more explicitly defined spatio-temporal peak to clarify that only the temporal scale is varied while spatial scale is fixed. The inter-scale saliency measure is defined for a fixed spatial scale:

$$\mathcal{W}_D([\hat{s}_s\ \hat{s}_t]^\mathsf{T}, \mathbf{x}) = s_t \sum_{d \in D} |b(d, s_s, s_t, \mathbf{x}) - b(d, s_s, s_t - \Delta s_t, \mathbf{x})| \tag{5.6}$$

For clarity, the subscript t has been added to $\mathcal{Y}_D(\hat{s}, \mathbf{x})$ to differentiate between spatial saliency ($\mathcal{Y}_D(\hat{s}_s, \mathbf{x})_s$) and temporal saliency. Now a new spatio-temporal saliency term is defined as the weighted sum of the spatial and temporal saliency. If the previous feature extraction processes had not been carried out beforehand, this effectively biases the selected regions to either spatially salient, temporally salient, or spatio-temporally salient regions. However, with initial processing, the measure becomes an effective measure of the spatio-temporal saliency of just moving regions of the image sequence so the spatio-temporal saliency at the spatio-temporal scale $\hat{s} = [\hat{s_s}\ \hat{s_t}]^\mathsf{T}$ is defined as:

$$\mathbf{Y}_D(\hat{s}, \mathbf{x}) = \mathcal{Y}_D(\hat{s}, \mathbf{x})_t + \mathcal{Y}_D(\hat{s}_s, \mathbf{x})_s\ \xi \tag{5.7}$$

where ξ is a weight attributed to the spatial saliency of the homogeneous region. Note that to reduce temporal scale ambiguity, the cylindrical kernel is only extended backwards in time whereas the spatial scales are expanded radially in space.

In the previous chapters, we have defined saliency in terms of the top percentage of salient regions. However, here, we are searching for regions which are both spatially and temporally salient over a wide area. Therefore, something which is very salient in one region may not be so over the entire image frame but it should still be considered given its local context. Furthermore, we have already shown in Figure 4.8, the effect of depth perspective on the typical temporally salient values. Therefore it is likely that regions of the scene which are further from the camera should be treated within its own context too. Rather than rank all saliency values globally and extract those that have the highest value, the spatio-temporal algorithm has an extra stage that selects salient volumes based on local spatial as well as global

(a) (b)

Figure 5.3: Two example frames of correctly detect spatio-temporally salient regions. Regions are highlighted by lighter shading. (a) Moving trees and fountain being ignored whilst detecting salient people (b)Person (left) detected despite occlusion by tree.

variations in saliency. Using a block-based approach, all saliency values within an equally subdivided block of the whole frame are accumulated to find the local median saliency value $\tilde{\mathbf{Y}}_{block}$. A mean meso-level saliency value is also accumulated from the whole image frame, and finally a global sampled median is accumulated from all the saliency values measured so far. Using this local, meso and global level measure of saliency, we can decide whether the region is salient enough to be considered. The term *meso* is used here to describe a mid-level context where saliency values within a local block (from evenly subdividing the image frame into a grid) are accumulated over time and salient regions are selected based on the distribution of the saliency values within these blocks.

Figure 5.3 shows some results using the spatio-temporal saliency method. The algorithm was able to detect salient people while ignoring motion from the trees and fountain in most cases, as well as detecting people that were partially occluded behind moving tree branches. Though it is difficult to see from the figure, the responses shown for each detected person are the result of many salient responses cluttered around the spatial neighbourhood of the detected person. Therefore, the person in the bottom left of Figure 5.3(a) appears to have two responses at quite considerable spatial displacement from each other. This particular scene also exhibited severe variations in the background at non-periodic intervals, which was mostly

(a) (b)

Figure 5.4: Two example frames where the background is selected. Regions are highlighted by lighter shading. (a) Moving trees detected while fountain is ignored. (b) Leaf motion from tree and also the fountain is detected due to a gust of wind and also a change in the height of the water of the fountain.

caused by sudden gusts of wind. There was also motion from a fountain in the middle of the courtyard that switched itself on and off periodically. These regions both yielded spatio-temporally salient responses when the background behaviour become less predictable. Two examples of this are shown in Figure 5.4. The problem with working such scenes, however, is that it becomes much more difficult to detect more detailed body behaviour from each person. Therefore, it is necessary to address the problem in terms of finding salient human forms at close range.

5.2.2 Humans at close range

In this subsection, focus turns to extracting salient human forms at close range. While extracting salient human forms at close range may seem like an easier problem, in some ways more challenging. At higher resolutions, background clutter may also be represented at higher resolution and therefore the foreground/background discrimination task becomes more difficult. Furthermore, it is much easier to discriminate between different speeds of motion, particularly from different limbs of the same person. So a person can produce a wide range of different temporally salient values, making it more difficult to attribute very different levels of saliency to the same moving object. If we wish to identify limbs of the body for natural

gesture recognition, it is important that limbs that move slowly are also selected.

Let us start by re-addressing how to extract temporally salient features from the scene. At this level, we can discard the temporal differencing stage employed in the previous subsection since the resolution of each person is much higher and they are proportionately larger with respect to the rest of the image frame, thus accommodating more severe sub-sampling of the frame. This gives us a more consistent approach to the overall method for extracting temporally salient motion. However, since we cannot perform spatial saliency on the temporal difference image, this leads to more possibilities of spatially salient ambiguity if the background is particularly cluttered. That is, something which is more spatially salient in one frame is not necessarily more temporally salient. However, we would wish that if two regions are similarly temporally salient but one is much more spatially salient, then we would prefer the more spatially salient region to have a higher saliency value. So here, the problem of quantifying both spatial and temporal saliency in a more globally discriminative way is addressed.

So far, the temporal saliency algorithm and the modifications described in the previous subsection could go further towards extracting salient spatially homogeneous and temporally salient regions. The first and perhaps more obvious problem with both the temporal saliency algorithm and Kadir and Brady's spatial saliency algorithm is that the measure of saliency relies solely on the characteristics within a small neighbourhood of scales of the inter-scale saliency measure $\mathcal{W}_D$, as defined in Equation 3.8, and entropy $\mathcal{H}_D$ around the spatial or temporal scale around which the entropy peaks. However, Figures 3.2 and 3.7 in Chapter 3 show spatial and spatio-temporal entropy-scale characteristics where it is clear that the entire curve provides a lot of information about the spatial or temporal dynamics of a particular region of imagery data, which is not being exploited. This can be rectified if the saliency measure is weighted by the variance of the entropy over all scales. However, the variance alone may be too sensitive to the change in entropy at lower scales. This becomes particularly problematic if the number of scales that are considered is small and the 'tail' of the entropy-scale characteristic is not fully formed. Of course we could just increase the number of scales but this is at the considerable cost of computation. Therefore, an option which is less sensitive

to fewer scales would be to replace the mean of the entropy and the variance with a median calculation.

We can define the median of variances in entropy over spatial scale, $\sigma_s(\mathbf{x})$ as the sample median of the difference between each entropy value and the sample median of the reordered entropy values over spatial scale $\tilde{\mathcal{H}}_D(\mathbf{x})$. Therefore, the spatial saliency measure is redefined as:

$$\mathcal{Y}_D(\hat{s_s}, \mathbf{x})_{weighted} = \mathcal{H}_D(\hat{s_s}, \mathbf{x})\ \mathcal{W}_{D_{peak}}\ \sigma_s \tag{5.8}$$

The median of variances over temporal scale, $\sigma_t(\mathbf{x})$ is similarly defined except the reordered entropy values are taken over all temporal but a single fixed spatial scale. Using this modification, the temporal saliency measure from Equation (5.5) is re-defined as

$$\mathcal{Y}_D(\hat{s}, \mathbf{x})_{weighted} = \mathcal{H}_D(\hat{s}, \mathbf{x})\ \mathcal{W}_{D_{peak}}\ \sigma_t \tag{5.9}$$

where $\hat{s}$ is again defined as the spatio-temporal peak in entropy, $\mathbf{x}$ specifies the spatio-temporal location around which the measure is calculated, and $\mathcal{W}_{D_{peak}}$ is the inter-scale saliency measure defined in Equation (3.8).

A further note to add here is that the longer the temporal kernel (or context is), the more difficult it becomes to determine the exact spatial location of object. However, if we consider fewer temporal scales where spatial alignment is more certain, we are faced with reduced certainty about how temporally salient the region actually is in a wider temporal context. To overcome this, we can make a further constraint on the algorithm by ensuring that while a larger number of temporal scales are considered, selection of salient temporal scales are considered for a fraction of this. Using this approach, it is possible to acquire a meso-level temporal context about the salient region since the median of variances can be acquired from the full range of temporal scales while scale selection is limited to scales which are temporally closer to the current frame of interest.

The spatial saliency algorithm is further modified to bias it towards salient regions located at the centre of a homogeneous rather than at the edge of it. (Kadir et al., 2004) suggested that by weighting the accumulated histogram by a windowing function, regions located at

the centre of a homogeneous region were more likely to have a higher saliency value than those at the edges of the region since the change in intensity distribution around the scales at which the entropy peaks would need to be pronounced in order to yield a high saliency value. Therefore, for each spatial location the likelihood of a particular pixel intensity is weighted by the following function:

$$SW(z) = \frac{1}{1 + \left(\frac{z}{s_s}\right)^{\nu}} \tag{5.10}$$

where z is measured as the distance of the pixel from the centre of the sampling kernel so that in the generalised case of an elliptical kernel with no orientation and a ratio of ρ between the major and minor axis, the distance is defined by Kadir (2002) as:

$$z = \sqrt{\frac{x'}{\sqrt{\rho}} + (y'\sqrt{r})^2} \tag{5.11}$$

where x' and y' are the horizontal and vertical distance respectively, of the pixel from the centre of the kernel. Using this windowing function, Kadir and Brady found that it was possible to make the spatial saliency measure more sensitive to blob-like structures rather than just edges. However, in the case of more cluttered scenes, using this weighting function is not enough to produce a suitably discriminative ordering of salient features. This is particularly the case if we are comparing a homogeneous highly textured region with one with a tighter intensity distribution since the textured region is more likely to have a lower inter-scale saliency value compared to that of a region with a tighter intensity distribution. Therefore a salient region centred around the boundary of the region with a tighter intensity distribution would be considered more salient than a region, which is located closer to the centre of a more homogeneous textured region.

Of course, we cannot make arbitrary decisions about which region is more important without more contextual information but even if we consider these two regions over time, ambiguity still arises around the boundary of a moving object against cluttered background. As shown in Figure 5.1 of the previous Section 5.1 the temporal saliency algorithm selected a temporally salient region but it was not at all spatially salient at the most recent frame

of the kernel (though it might have been a few frames later). This was rectified to some extent, in the previous subsection where only seed locations which were spatially salient were considered. However, the edge of an object can still be considered spatially salient under the right contextual conditions and so a more robust way of determining temporally salient and spatially homogeneous regions is required. We can treat the problem as one of spatial alignment of the spatio-temporal kernel. Since we wish to determine whether a particular homogeneous region of interest is spatio-temporally salient, it is important that the spatial saliency of the current frame of the kernel has a higher value than that of the frame that corresponds to the salient temporal scale. By imposing this constraint, the receptive field becomes much less sensitive to salient changes of a spatio-temporal region, which changes from background to foreground rather than vice versa.

Another limitation of the temporal saliency algorithm which has been used so far is that it has been insensitive to colour changes. If we were to consider a three-dimensional colour space, composed of the red, blue and green channels, the complexity of the algorithm would be increased greatly so mapping this to a two-dimensional colour space would be preferable. In addition, increasing the dimensionality of the feature space would lead to sparsity problems or a reduction in sparsity at the expense of lower bin resolution. Therefore, the algorithm was modified to work in a two-dimensional colour space defined by the hue and saturation from RGB values. So for any spatio-temporal $\mathbf{x} = [x\ y\ t]^\mathsf{T}$ or spatial location $\mathbf{x} = [x\ y]^\mathsf{T}$, with a RGB colour vector $[\mathcal{R}\ \mathcal{G}\ \mathcal{B}]^\mathsf{T}$ representing the value of the red, green and blue channel respectively, we can find their corresponding hue, saturation and value colour vector $[\mathcal{H}\ \mathcal{S}\ \mathcal{V}]^\mathsf{T}$ using the equations below.

$$\mathcal{H} = \begin{cases} \text{undefined} & \text{, if } min(\mathcal{R},\mathcal{G},\mathcal{B}) = max(\mathcal{R},\mathcal{G},\mathcal{B}) \\ 60\frac{\mathcal{G}-\mathcal{B}}{max(\mathcal{R},\mathcal{G},\mathcal{B})-min(\mathcal{R},\mathcal{G},\mathcal{B})} & \text{, if } max(\mathcal{R},\mathcal{G},\mathcal{B}) = \mathcal{R} \text{ and } \mathcal{G} \geq \mathcal{B} \\ 60\frac{\mathcal{G}-\mathcal{B}}{max(\mathcal{R},\mathcal{G},\mathcal{B})-min(\mathcal{R},\mathcal{G},\mathcal{B})} + 360 & \text{, if } max(\mathcal{R},\mathcal{G},\mathcal{B}) = \mathcal{R} \text{ and } \mathcal{G} < \mathcal{B} \\ 60\frac{\mathcal{B}-\mathcal{R}}{max(\mathcal{R},\mathcal{G},\mathcal{B})-min(\mathcal{R},\mathcal{G},\mathcal{B})} + 120 & \text{, if } max(\mathcal{R},\mathcal{G},\mathcal{B}) = \mathcal{G} \\ 60\frac{\mathcal{R}-\mathcal{G}}{max(\mathcal{R},\mathcal{G},\mathcal{B})-min(\mathcal{R},\mathcal{G},\mathcal{B})} + 240 & \text{, if } max(\mathcal{R},\mathcal{G},\mathcal{B}) = \mathcal{B} \end{cases}$$

$$\mathcal{S} = \begin{cases} 0 & \text{, if } max(\mathcal{R},\mathcal{G},\mathcal{B}) = 0 \\ 1 - \frac{min(\mathcal{R},\mathcal{G},\mathcal{B})}{max(\mathcal{R},\mathcal{G},\mathcal{B})} & \text{, otherwise} \end{cases}$$

$$\mathcal{V} = max(\mathcal{R},\mathcal{G},\mathcal{B}) \tag{5.12}$$

It is important to note at this stage that we have not made an in-depth study of different colour spaces since the aim of using colour is just to maximise the discriminance between different shades of colour while still only introducing one extra dimension to the histogram representation. More information about perceptually uniform colour spaces can be found in (Ma and Zhang, 2003; Shafarenko et al., 1997).

Although we are now working in two-dimensional colour space, the colour distribution can still be treated as a concatenated one-dimensional histogram since the inter-scale saliency measure $\mathcal{W}$ is only a measure of the sum of absolute difference. Intuitively, it may be beneficial to treat hue-saturation distributions as a mixture of Gaussians and then calculate the distance between these Gaussian distributed forms over consecutive scales. However, this would greatly increasing the computational complexity since the mixture of Gaussians of the colour distributions would have to be re-estimated at every scale. Further complications of model order selection would also be introduced since it would be impossible to estimate a GMM with empty sets and it would no longer make sense to quantify the data in terms of entropy. In any case, treating all the bins of the hue-saturation distributions independently gave sufficiently improved results compared to using just grey levels. It is important to note here that while introducing colour discriminance is beneficial, it does cause some issues to do with the entropy values since we are now dealing with a much larger feature space than before. That is, in cluttered scenes, even if a homogeneous region of interest exists, if the background around it is much more cluttered, the likelihood of the colour distribution of the background region dominating the colour distribution of the salient region of interest is much less at each scale increase compared to if the salient region lies on a uniform background. This means that the salient scale of a region of interest will be much larger if it is surrounded by cluttered background rather than a uniform background, leading to further spatial ambiguity. In summary, the modified spatio-temporal saliency algorithm is described in Algorithm 1.

tb **for** *each pixel of the current image frame* **do**
Calculate the hue-saturation distribution, $b_{\mathcal{HS}}$ for all spatio-temporal scales that are considered;
for *each* s_t **do**
for *each* s_s **do**
Calculate the entropy $\mathcal{H}_D(\mathbf{x})$ of the hue-saturation distribution ;
end
end
$\hat{\mathbf{s}}_s$: Calculate the peaks in $\mathcal{H}_D(\mathbf{x})$ across spatial scales where $s_t = 0$;
Calculate the inter-scale saliency $\mathcal{W}_D$;
for *Each peak in the feature vector* $\hat{\mathbf{s}}_s$ **do**
$\hat{\mathbf{s}}_t$: Calculate the peaks in $\mathcal{H}_D(\mathbf{x})$ over temporal scales where $s_s = \hat{s}_s$;
if *Peaks in* $\hat{\mathbf{s}}_t$ *exist,* **then**
Calculate the spatial saliency, $\mathcal{Y}_D([\hat{s}_s\ 0]^\mathsf{T}, \mathbf{x})$ and at $s_t = 0$;
if *there is a peak over spatial scales in the* $\mathcal{H}_D(\mathbf{x})$ *at* $s_t = \hat{s}_t$ *and* $s_s = \hat{s}_s$ **then**
Calculate the spatial saliency, $\mathcal{Y}_D([\hat{s}_s\ \hat{s}_t]^\mathsf{T}, \mathbf{x})$ at $s_t = \hat{s}_t$,$s_s = \hat{s}_s$;
if $\mathcal{Y}_D([\hat{s}_s\ \hat{s}_t]^\mathsf{T}, \mathbf{x}) > \mathcal{Y}_D([\hat{s}_s\ 0]^\mathsf{T}, \mathbf{x})$ **then**
Discard this peak;
end
end
end
end
for *All the spatio-temporally salient peaks that are left,* **do**
Calculate the inter-scale saliency $\mathcal{W}_D$ over temporal scales at the current salient spatial scale. Calculate the temporal saliency at this point.
end
end

Algorithm 1: Description of the spatio-temporal saliency algorithm.

Once the temporally salient regions have been found, a saliency map for each frame is obtained by accumulating the saliency values over time depending on the salient spatio-temporal scale of the volume. This produces a frame by frame accumulated temporal saliency map of the sequence. In order to select the salient human forms from this, we apply the elliptical spatial saliency algorithm on the temporal saliency map to find temporally salient elliptical forms. Then it is necessary to group them to find the salient human forms. The membership of an ellipse in a group is determined by whether it overlaps spatially with any other ellipses within the group. Given two ellipses i and j, as shown in Figure 5.5(a), the line l intersects both of their centroids, $\mathbf{x}_i$ and $\mathbf{x}_j$, and d_{ij} is the distance between them. $\mathbf{o}_{i1}$ and $\mathbf{o}_{i2}$ are points on the perimeter of ellipse i that intersects line l and Δ_i is the distance from $\mathbf{x}_i$ to one of these points (or half the distance between $\mathbf{o}_{i1}$ and $\mathbf{o}_{i2}$). These dimensions are used to measure a degree of overlap between two ellipses. The overlap measure is computed using the distance Δ_i and Δ_j which are weighted against the relative scales of both ellipses and normalised by the distance between their centroids, d_{ij}. However, this is an approximation and in the special case where one ellipse is much smaller than the other and is completely overlapped by the other, the degree of overlap is set to an overlap threshold O_{Thres}. Therefore the proximity measure is defined as:

$$O_\gamma(i,j) = \begin{cases} O_{Thres} & \text{if } \left((sq(\mathbf{o}_{i1}) - sq(\mathbf{x}_j)) \begin{bmatrix} 1/\sigma_{j1}^2 \\ 1/\sigma_{j2}^2 \end{bmatrix} < 1 \right) \wedge \left((\mathbf{o}_{i2}^2 - \mathbf{x}_j^2) \begin{bmatrix} 1/\sigma_{j1}^2 \\ 1/\sigma_{j2}^2 \end{bmatrix} < 1 \right) \\ \frac{\left(\Delta_i \frac{s_i}{s_j} + \Delta_j \frac{s_j}{s_i} \right)}{d_{i,j}} & \text{otherwise} \end{cases} \tag{5.13}$$

where $sq(\mathbf{x}) = [x\ y] \begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix}$, σ_1 and σ_2 are the distances between the centroid of the ellipse and the its edge along the minor and major axes respectively. The two terms on the top row and right hand side of the equation check that the perimeter point $\mathbf{o}_{i1}$ is within ellipse j and vice versa.

The ellipses are grouped using the following algorithm:

Once the ellipses have been grouped, a bounding box generated for each group according to

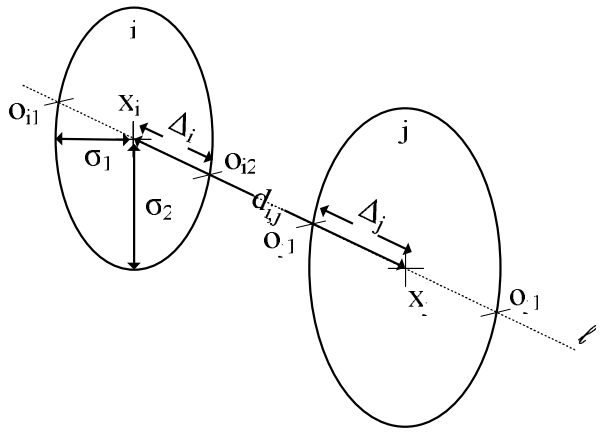


Figure 5.5: Measuring overlaps between two ellipse.

for *All the ellipses in the current frame* **do**
  if *this is the first ellipse in the list which has not been grouped* **then**
    Make a new group with this ellipse as the seed;
  end
  while *Ellipses can still be added to the latest group* **do**
    if *this ellipse overlaps any of the ellipses in the group* **then**
      add this ellipse to the group;
      mark the ellipse as already being grouped ;
    end
  end
end

Algorithm 2: Algorithm for grouping the ellipses

the extremities of each group of ellipses. Each box is then divided into 3 equal vertical parts, which represent a meso-level of context within which less salient ellipses can be discarded depending on whether the saliency of each ellipse is above a percentage of the maximum saliency value for the ellipses in that particular part of the human bounding box. Only a proportion of the person should be considered since the temporal saliency from the motion of the feet is likely to be greater than that of the upper half of the body when a person is walking and so salient ellipse selection should be based on the relative saliency within each of these parts.

5.3 Classifying human-human interaction

5.3.1 A pose-distribution model

In the previous section, a method for finding salient human forms at close range was formulated. If we wish to classify the actions of a particular person or pair of people, then it is necessary to represent their behaviour in some way. Using the spatial saliency algorithm it is intuitive to find the orientation of homogeneous salient regions of interest. By extracting the orientations of regions of interest rather than just edges, we are applying a higher semantic meaning to the extracted features, which may represent limbs of the body.

Once each bounding box, representing the spatial position of a person, has been trimmed of any less salient ellipses within the context of each part, the temporal saliency described in this section is applied again but at a finer spatial grid where spatially salient seed locations are identified before the temporal saliency algorithm is used. Again, temporally salient regions found using the finer scales are ranked according to the meso-level context of salient regions for each part of each person individually. Each part of the person is one of three equal vertically distributed regions of the bounding box of each person.

In keeping with the idea of using entropy to measure the homogeneity of a spatial region, the affine version of the scale saliency algorithm was used to identify different limbs of the body. However, it was found that the method was unstable at low resolution and computationally intensive. It was found that the method ran faster and was more robust with binary

silhouettes of the human form. However, this required a precise silhouette of the human body and it only provided shape information about the silhouette of a person. A better method, which is both faster and able to deal with limbs which spatially overlap other parts of the body involves using central moments to identify the elongation and orientation of a particular blob.

Usually moments are calculated by firstly determining a region in which to calculate its moments based on the membership of a binary mask. However, in this case, we don't have a binary image and so another type of membership function is defined. Since each temporally salient region also has an associated spatially salient scale, we can exploit this in order to build a membership function of the salient region of interest based on the mean of its colour distribution. Therefore at every location $\mathbf{x}_c$ [1] where a spatially salient scale, $\hat{s}_s$, exists, each element of the circular membership function associated with the salient region is defined as the absolute distance between the colour value of each pixel in the region in its corresponding hue-saturation space and the mean hue and saturation of the whole circular region.

$$m(\mathbf{x})_{\mathbf{x}_c} = |[\mathcal{H}\ \mathcal{V}]^{\mathsf{T}} - [\overline{\mathcal{H}}\ \overline{\mathcal{V}}]^{\mathsf{T}}| \tag{5.14}$$

where $\mathbf{x}_c$ denotes the point around which the salient scale was found and $\overline{\mathcal{H}}$ and $\overline{\mathcal{V}}$ are the mean hue and saturation values for the salient region. The mean of the hue and saturation are calculated by firstly mapping the values into the imaginary space. This was necessary since we must find the median of the colour distribution over a circular surface in hue-saturation space. So the hue and saturation can be seen as polar coordinates that must be mapped into Cartesian coordinates before calculating the mean. Let $Cart_{\mathrm{x}}$ and $Cart_{\mathrm{y}}$ be the two new components in the Cartesian domain which we wish to calculate mean of the distribution so

$$Cart_x = \mathcal{S}\cos\mathcal{H} \tag{5.15}$$

$$Cart_y = \mathcal{S}\sin\mathcal{H} \tag{5.16}$$

[1] The subscript c has been used here to distinguish the actual centroid of a salient region and the standard notation used in this thesis, to represent a spatio-temporal location.

and their corresponding mean values for the colour distribution b_{col} is defined as

$$\overline{Cart_x} = \frac{1}{\sum_{\mathrm{h}=1}^{2\pi}\sum_{\mathrm{a}=0}^{1} b_{col}(\mathcal{H}_\mathrm{h}, \mathcal{S}_\mathrm{a}, s_s, \mathbf{x})} \sum_{\mathrm{h}=1}^{2\pi}\sum_{\mathrm{a}=0}^{1} b_{col}(\mathcal{H}_\mathrm{h}, \mathcal{S}_\mathrm{a}, s_s, \mathbf{x}) Cart_x \tag{5.17}$$

$$\overline{Cart_y} = \frac{1}{\sum_{\mathrm{h}=1}^{2\pi}\sum_{\mathrm{a}=0}^{1} b_{col}(\mathcal{H}_\mathrm{h}, \mathcal{S}_\mathrm{a}, s_s, \mathbf{x})} \sum_{\mathrm{h}=1}^{2\pi}\sum_{\mathrm{a}=0}^{1} b_{col}(\mathcal{H}_\mathrm{h}, \mathcal{S}_\mathrm{a}, s_s, \mathbf{x}) Cart_y \tag{5.18}$$

We can then obtain the original mean hue and saturation values as follows:

$$\overline{\mathcal{H}} = tan^{-1}\left(\frac{Cart_y}{Cart_x}\right) \tag{5.19}$$

$$\overline{\mathcal{S}} = \begin{cases} 0 & \text{if } \overline{\mathcal{H}} = 0 \\ \frac{\overline{Cart_y}}{sin(\overline{\mathcal{H}})} & \text{otherwise} \end{cases} \tag{5.20}$$

The central moments can now be used to calculate the elongation and orientation based on the membership of each pixel in the salient region. The centroid of each region is calculated using Equations (7.2) and (7.3) except now rather than weighting by the binary image map, we weight by the distance in hue-saturation space between the current pixel colour and the mean colour of that region. The orientation may be defined as the angle of axis of the minimised moment of inertia which can be expressed in terms of the second order central moments as:

$$\theta(s_s, \mathbf{x}) = \frac{1}{2} tan^{-1}\left[\frac{2\mu_{11}}{\mu_{20} - \mu_{02}}\right] \tag{5.21}$$

where θ is expressed with respect to the x axis and μ_{ij} can be expressed using the following equation:

$$\mu_{ij} = \sum_x \sum_y (x - \overline{x})^i (y - \overline{y})^j \, m_{xy} \tag{5.22}$$

where m is the membership function or distance of each pixel within the salient region with respect to the mean hue-saturation value of that neighbourhood of pixels. The elongation or eccentricity of the region is defined using Equation 7.6. Using the elongation, scale and orientation of salient regions of interest, it is possible to accumulate a co-occurrence distribution of these features for each part of the bounding box of each human form. If these are

accumulated for each time frame, a feature vector is created that describes the approximate pose information of each person. This pose co-occurrence histogram pp is defined as:

$$pp_{\mathrm{c}} = \sum_{\mathrm{c}} \delta[bin(\theta(\hat{s}, \mathbf{x})) - \mathrm{i}]\delta[bin(Elong(\hat{s}, \mathbf{x})) - \mathrm{j}]\delta[bin(s_s(\mathbf{x})) - k] \tag{5.23}$$

where c indexes all salient regions within the current part of the human bounding box and the δ function is simply the standard function which has the value of 1 at 0 and is 0 elsewhere. A feature vector for each person at each frame is created by concatenating the co-occurrence histogram for each part into a single vector so that a pair of vectors is created for each action.

5.3.2 Creating feature trajectories from the pose distribution model

Once each bounding box has been generated for each group of ellipses, trajectories are formed based on whether the bounding box overlaps with the previous boxes in the trajectory. In the first instant, trajectories are not formed until two spatially separated bounding boxes are detected. At each frame, the width of each bounding box is kept so that when only one box is detected, the amount of overlap of each box is calculated based on the mean width of each. When the overlap becomes too great, the algorithm is stopped.

5.3.3 Canonical Correlation Analysis for modeling temporally correlated behaviour

In the previous chapter, accumulated co-occurrences were used to represent temporally correlated behaviour between salient features. In this case, the within-person pose distributions are modeled using a co-occurrence distribution but there is nothing to represent the inter-person temporal binding.

Canonical correlation analysis (CCA) (Hotelling, 1936) is a useful method of representing the relationship between two multi-dimensional variables since it finds two bases in which both vectors are optimally correlated. Using this technique, the two feature co-occurrence vectors formed from Equation (5.23) are correlated together where a sliding window over the whole sequence is used such that each frame of the window is treated as a data point in the

CCA calculation. Therefore for the whole sequence, a set of canonical correlations between the pose co-occurrence histograms of the two people is formed.

For the following descriptions of the algorithm, notation has in part been derived from (Shan et al., 2007). Formally, CCA is defined in terms of two sets of random variables $\mathbf{q} \in R^m$ and $\mathbf{r} \in R^n$ where the goal is to find a pair of basis vectors, $\mathbf{w}_q$ and $\mathbf{w}_r$ respectively, which maximises the correlation between the projected versions of $q = \mathbf{w}_q^T\mathbf{q}$ and $r = \mathbf{w}_r^T\mathbf{r}$ in canonical space. In other words, the correlation matrix between the variables is diagonal where the variables on the diagonal are maximised.

For the case where only one pair of basis vectors are needed, i.e. the one corresponding to the largest canonical correlation, then we are trying to maximise the correlation between the projections q and r respectively, as the canonical variates $q = \mathbf{w}_\mathbf{q}^T\mathbf{q}$ and r $= \mathbf{w}_r^T\mathbf{r}$. Therefore ρ must be maximised:

$$\begin{aligned}\rho &= \frac{E[qr]}{\sqrt{E[q^2]E[r^2]}} = \frac{E[\mathbf{w}_q^T qr^T \mathbf{w}_r]}{\sqrt{E[\mathbf{w}_q^T qq^T \mathbf{w}_q]E[\mathbf{w}_r^T rr^T \mathbf{w}_r]}} \\ &= \frac{\mathbf{w}_q^T \mathbf{C}_{qr}\mathbf{w}_r}{\sqrt{\mathbf{w}_q^T \mathbf{C}_{qq}\mathbf{w}_q \mathbf{w}_r^T \mathbf{C}_{rr}\mathbf{w}_r}}\end{aligned} \tag{5.24}$$

where $\mathbf{C}_{qq}$ and $\mathbf{C}_{rr}$ are the within set covariance matrices of $\mathbf{q}$ and $\mathbf{r}$ respectively and $\mathbf{C}_{qr} = \mathbf{C}_{qr}^T$ is the between set covariance matrix.

The canonical correlations between $\mathbf{q}$ and $\mathbf{r}$ are found by solving the following equations:

$$\begin{cases} \mathbf{C}_{qq}^{-1}\mathbf{C}_{qr}\mathbf{C}_{rr}^{-1}\mathbf{C}_{rq}\mathbf{w}_r = \rho^2\mathbf{w}_q \\ \mathbf{C}_{rr}^{-1}\mathbf{C}_{rq}\mathbf{C}_{qq}^{-1}\mathbf{C}_{qr}\mathbf{w}_q = \rho^2\mathbf{w}_r \end{cases} \tag{5.25}$$

such that the eigenvalues, ρ are the canonical correlations and eigenvectors $\mathbf{w}_q$ and $\mathbf{w}_r$ are their corresponding basis vectors between $\mathbf{q}$ and $\mathbf{r}$.

In the case of our feature trajectories, for each sequence, a pair of feature trajectories $\mathbf{Q}$ and $\mathbf{R}$ is created to represent the co-occurrence of the elongation and orientation of salient regions for each person using Equation 5.23. In order to represent the temporally correlated behaviour of the two people, CCA is computed between the two feature vectors of the two

people for a sliding time window so that for each window, a set of canonical factor pairs exists. Let us modify the notation slightly at this point so that the set of features in one sliding time window is represented by $\mathbf{q}' = \{\mathbf{q}_1, \ldots, \mathbf{q}_\mathrm{m}\}$ and $\mathbf{r}' = \{\mathbf{r}_1, \ldots, \mathbf{r}_\mathrm{m}\}$ where now m is the window size. $\mathbf{Q} = \{\mathbf{q}'_1, \ldots, \mathbf{q}'_\mathrm{n}\}$ and $\mathbf{R} = \{\mathbf{r}'_1, \ldots, \mathbf{r}'_\mathrm{n}\}$ each contain sets of feature vectors for each sliding time window where n is the number of sliding windows in the sequence.

We can map each sample point in the sliding window of both $\mathbf{q_i}$ and $\mathbf{r_i}$ respectively, into the reduced correlation space of $\mathbf{q}'$ and $\mathbf{r}'$, using their leading factor pairs $\hat{\mathbf{w}}_q$ and $\hat{\mathbf{w}}_r$ respectively.

$$\tilde{q}_\mathrm{i} = \hat{\mathbf{w}}_q^T \cdot q_\mathrm{i} \tag{5.26}$$

$$\tilde{r}_\mathrm{i} = \hat{\mathbf{w}}_r^T \cdot r_\mathrm{i} \tag{5.27}$$

From these projections, we can synthesise new versions of the feature vector q_{synth} or r_{synth} of person r from person q respectively using linear regression. For clearer notation, the i has been dropped here.

$$q_{synth} = \mathbf{U}_{\mathbf{q}'}^T \ \tilde{r} \tag{5.28}$$

$$r_{synth} = \mathbf{U}_{\mathbf{r}'}^T \ \tilde{q} \tag{5.29}$$

where q_{synth} and r_{synth} are the corresponding synthesised feature vectors from the projections of r_i and q_i respectively into the reduced correlation space. The regression matrices, which are used to reconstruct these synthesised vectors for this particular time window are defined

$$\mathbf{U}_{\mathbf{q}'} = (\mathbf{r}'^T \hat{\mathbf{w}}_r)(\mathbf{q}'^T)^{-1} \tag{5.30}$$

$$\mathbf{U}_{\mathbf{r}'} = (\mathbf{q}'^T \hat{\mathbf{w}}_q)(\mathbf{r}'^T)^{-1} \tag{5.31}$$

For a whole sequence, a set of synthesised versions of the trajectories $\mathbf{Q}$ and $\mathbf{R}$ are created as $\mathbf{Q}_{synth}$ and $\mathbf{R}_{synth}$. However, rather than having one synthesised value for each sample point, since the sliding time window overlaps many of the same frames, it is possible to have more than one possible synthesis of the sample point, as shown in Figure 5.7. Choosing which is the most appropriate at this stage would be in appropriate since it would depend

on the synthesised versions of the previous sample points in the sequence. Therefore, the selection process can be moved to a later stage by integrating the different versions into the path minimisation steps of the Levenshtein distance algorithm.

5.3.4 Interpolated Levenshtein distance

In order to find the distance between two feature trajectories, trajectory warping within a local neighbourhood is used. The interpolated Levenshtein distance (Kruskal and Liberman, 1999) is employed because it is possible to estimate the distance between two trajectories even if their sample points are significantly temporally shifted since the distance between two trajectories is based on the closest distance between a sample point on one trajectory and an interpolated point on the other, as shown in Figure 5.6.

Given two trajectories $\mathbf{Q} = \{\mathbf{q_1} \dots \mathbf{q_m}\}$ and $\mathbf{R} = \{\mathbf{r_1} \dots \mathbf{r_n}\}$ with lengths m and n respectively, where both $\mathbf{q_i}$ and $\mathbf{r_i}$ both exist in the space of real numbers, we can calculate the minimum distance between a sample point one trajectory and a line segment on the other by calculating the projected interpolation between two sample points in $\mathbf{Q}$ that is required at sample point $\mathbf{r_i}$. So an interpolation point, using the same notation as Kruskal and Liberman (1999) is $(\mathbf{q_i}, \xi)$ so that $(\mathbf{q_i}, \xi)$ can be thought of as being the distance ξ away from $\mathbf{q_i}$ towards $\mathbf{q_{i+1}}$ so that $(\mathbf{q_i}, \xi)$ is defined as:

$$(\mathbf{q_i}, \xi) = (1 - \xi)\mathbf{q_i} + \xi\mathbf{q_{i+1}} \tag{5.32}$$

where a preliminary measure of the required interpolation ξ' is calculated relative to the distance between the sample point $\mathbf{q_i}$ and $\mathbf{r_i}$ and the distance between consecutive sample points in one of the trajectories.

$$\xi' = \frac{(\mathbf{q_i} - \mathbf{r_j}) \cdot (\mathbf{r_{j+1}} - \mathbf{r_j})}{(\mathbf{r_{j+1}} - \mathbf{r_j}) \cdot (\mathbf{r_{j+1}} - \mathbf{r_j})} \tag{5.33}$$

where $\cdot$ represents the scalar product of two vectors. In order to maintain some local bootstrapping, if the value of ξ' projects the interpolated point beyond $\mathbf{r_j}$ or $\mathbf{r_{j+1}}$, ξ is formulated

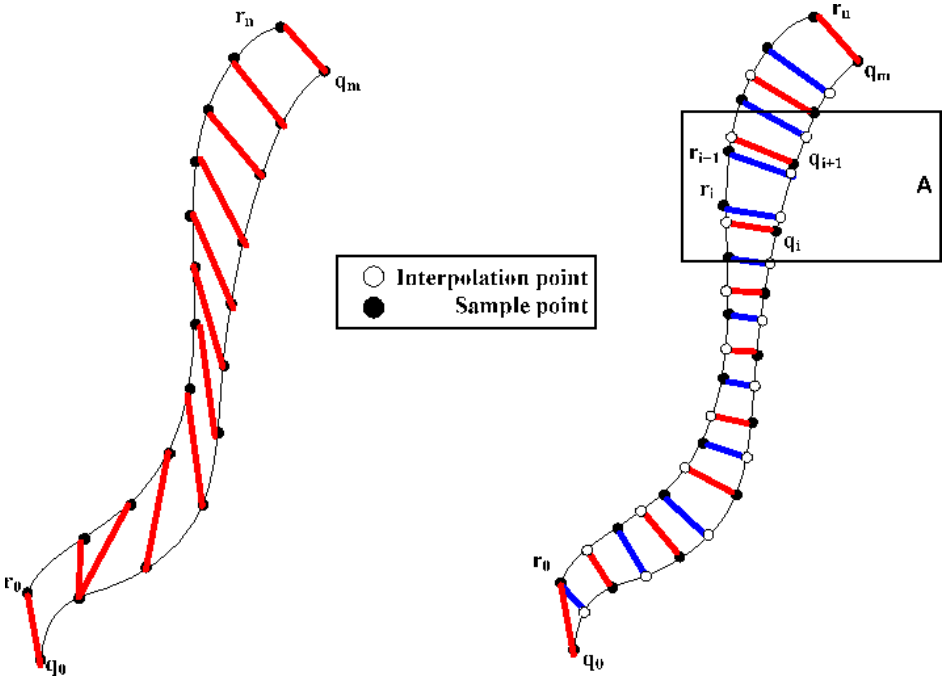


Figure 5.6: The distance between two trajectories using the traditional measure of distance between sample points (left) and using interpolation warping (right). The true distance between two trajectories can be severely misrepresented if the trajectories are misaligned temporally, as shown by the illustration on the left, even if their shape is very similar. However, using interpolation time warping, as shown on the right, it is possible to calculate the distance based on the nearest point of the trajectory from the sample point on the other. Sample points are represented by black dots whereas interpolation points are white circles. Part A on the right shows an example of where two interpolation points from one trajectory can correspond to the same line segment in the other. Using interpolation time warping allows these two points to belong to the same segment, and therefore leads to more flexible trajectory matching.

to ensure that the value is always between 0 and 1.

$$\xi = \begin{cases} 1 & \text{if } \xi' > 1 \\ 0 & \text{if } \xi' < 0 \\ \xi' & \text{otherwise} \end{cases} \tag{5.34}$$

Therefore the linkage cost between a sample point and its corresponding interpolated point

on the trajectory can be formulated using either of the following two relations:

$$w[\mathbf{r_i}, (\mathbf{q_j}, \xi)] = ||\mathbf{r_i} - [\mathbf{q_j} + \xi(\mathbf{q_{j+1}} - \mathbf{q_j}]||| \tag{5.35}$$

$$w[(\mathbf{r_i}, \xi), \mathbf{q_j}] = |||[\mathbf{r_i} + \xi(\mathbf{q_{j+1}} - \mathbf{q_j})] - \mathbf{q_j}|| \tag{5.36}$$

Then, in order to calculate the Levenshtein distance, a two-dimensional matrix is used to represent all the minimal accumulated distances between all points on both trajectories and the associated cost of each interpolated link. Using this, a path of minimal cost is found from start to end of both trajectories. Let $\mathbf{D}$ be the distance matrix so that

$$\mathbf{D}_{ij} = min \begin{cases} \mathbf{D}_{i-1,j} + & min_{0 \leq \xi \leq 1} w[\mathbf{r}_\mathrm{i}, (\mathbf{q}_\mathrm{j}, \xi)], \\ \mathbf{D}_{i,j-1} + & min_{0 \leq \xi \leq 1} w[(\mathbf{r}_\mathrm{i}, \xi), \mathbf{q}_\mathrm{j}]. \end{cases} \tag{5.37}$$

and the minimum cost of any set of trajectories derived using the interpolation functions described above is:

$$Levd(\mathbf{q_j}, \mathbf{r_i}) = \mathbf{D}_{m-1,n-1} + min \begin{cases} min_{0 \leq \xi \leq 1} w[\mathbf{r}_\mathrm{m}, (\mathbf{q}_\mathrm{n-1}, \xi)], \\ min_{0 \leq \xi \leq 1} w[(\mathbf{r}_\mathrm{m-1}, \xi), \mathbf{q}_\mathrm{n}]. \end{cases} \tag{5.38}$$

5.3.5 Reformulating the interpolated Levenshtein distance for multiple hypothetical trajectories

In order to compare pairs of test and model feature trajectories, it is necessary to use the dyadic relation between the two people to predict what the other feature vector will be for one person, given the other. In this case, a synthesised version of person 1 is created from person two or vice versa, using linear regression of the canonical factors from the training sequence where each frame of a sliding window represents one observation when the canonical factors are computed.

Calculating the Levenshtein distance is fairly straight forward for simple feature trajectories. However, since the canonical factors are calculated for a sliding time window, then projecting each frame of the feature vectors from the test sequence will lead to multiple can-

didates at each time frame, as shown in Figure 5.7. For classification, the synthesised versions

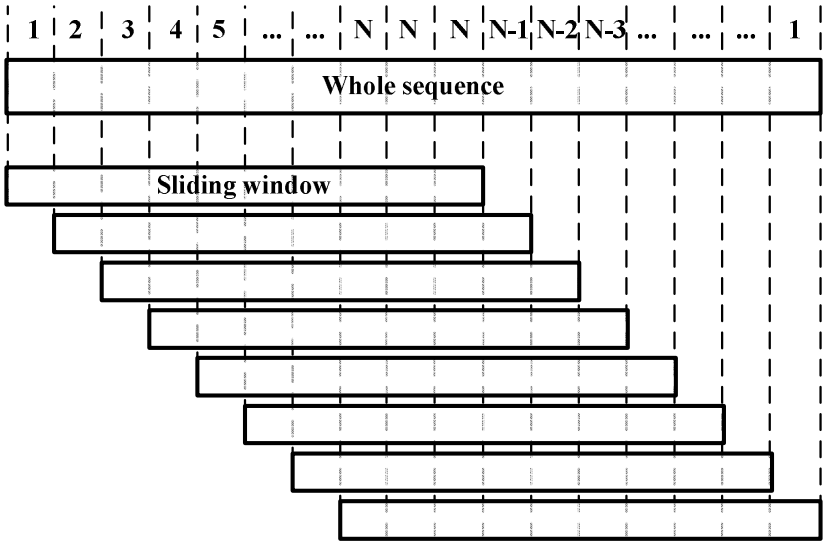


Figure 5.7: Figure showing the number of possible synthesised versions of the same frame for a sequence given the regression matrix which is created for each pair of sliding windows in the training data.

of the sequences are created using the regression matrices. Here, the leading correlation basis vectors created from the training data are used to synthesise person 1 from 2 or vice versa. Therefore, at each original test sample point, there are multiple possible synthesised versions of the same point where the number of possible synthesised options at $\mathbf{q}_\mathrm{j}$ can be formulated as

$$\Omega_\mathrm{j} = \begin{cases} j & \text{if}(j \leq \tau) \wedge (j \leq m - \tau + 1) \\ \Omega_{\mathrm{k}-1} - 1 & \text{if}(j > \tau) \wedge (j > m - \tau + 1) \\ \Omega_{\mathrm{k}-1} & \text{otherwise} \end{cases} \tag{5.39}$$

where $j \in \{1, \ldots, m\}$, and τ is the interval of the sliding window and k $=\in \{1, \ldots, max[\Omega]\}$. In addition, there are multiple possible interpolations of the same sample point. Depending on Ω_j, Equation (5.40) is reformulated so that for each sample point, j, there are Ω_j possible interpolations of the different synthesised versions of the data.

$$\xi'_\mathrm{k} = \frac{(\mathbf{q}_{\mathrm{i,k}} - \mathbf{r}_\mathrm{j}) \cdot (\mathbf{r}_{\mathrm{j}+1} - \mathbf{r}_\mathrm{j})}{(\mathbf{r}_{\mathrm{j}+1} - \mathbf{r}_\mathrm{j}) \cdot (\mathbf{r}_{\mathrm{j}+1} - \mathbf{r}_\mathrm{j})} \tag{5.40}$$

where ξ'_k is again capped to be between 0 and 1. So Equation (5.37) can then be reformulated

using the multiple hypothesis version of $\mathbf{q}$.

$$\mathbf{D}_{ij} = min \begin{cases} \mathbf{D}_{i-1,j} + & min_{\mathrm{k}} \left[min_{0 \leq \xi \leq 1} w[\mathbf{r}_{\mathrm{i}}, (\mathbf{q}_{\mathrm{j,k}}, \xi_{\mathrm{k}})] \right], \\ \mathbf{D}_{i,j-1} + & min_{\mathrm{k}} \left[min_{0 \leq \xi \leq 1} w[(\mathbf{r}_{\mathrm{i}}, \xi_{\mathrm{k}}), \mathbf{q}_{\mathrm{j,k}}] \right]. \end{cases} \tag{5.41}$$

A summary of the algorithm is shown in Figure 5.8.

5.3.6 Algorithmic optimisation of the spatial and temporal saliency algorithms

The temporal saliency algorithm was found to be quite slow in the previous chapter. There was much redundancy in the calculations which could have been eliminated. Computational redundancy has already been reduced, to some extent, through the modification of the temporal saliency algorithm described in Chapter 4 to a one-sided spatio-temporal sampling kernel, which halved the computation time and also led to responses that were less temporally ambiguous.

More serious computational gains can be made by firstly making the histogram calculations on-line so that at each new frame the intensity histogram vector of the current frame for each sub-sampled pixel location and scale is stored. Then, a second variable is introduced to store the accumulated intensity histogram for each sliding spatio-temporal kernel. At each frame, rather than re-accumulating the intensity histogram, the histogram can be accumulated or deccumulated according to the frame-based histograms.

5.3.7 Preliminary experiment on classifying interaction and non-interaction

The algorithm described above was applied to 3 different interaction or non-interaction scenarios, namely, shaking hands, waving and neither. For each action, two people walk into the scene, perform or don't perform an interaction and then walk off the scene. Examples of the 3 classes are shown in in Figure 5.9. Here the top row shows the shaking hands gestures, the second row shows the waving gesture and the last row shows the non-interaction state.

The results from estimating the poses is shown in Figure 5.10. Here, the original frame is shown, as well as the temporal saliency map and the resultant pose estimates.

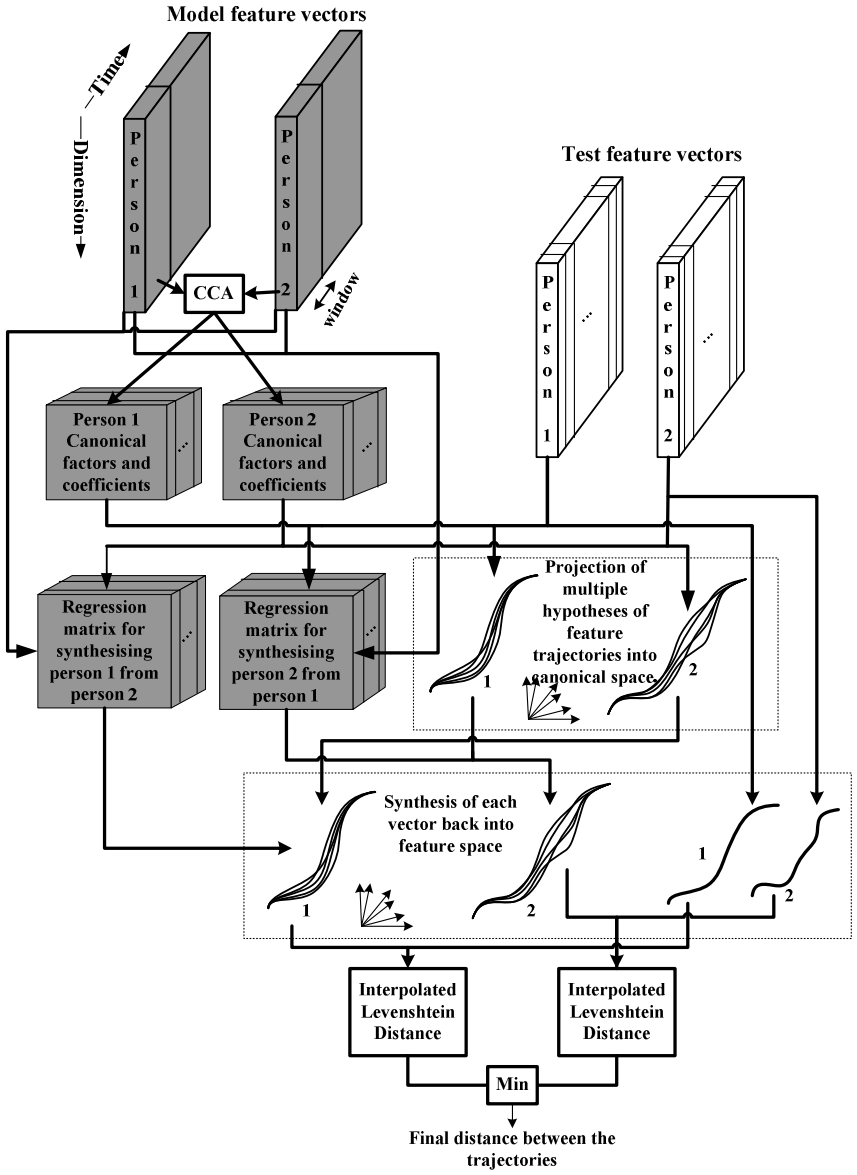


Figure 5.8: Summary of the human interaction classification system.

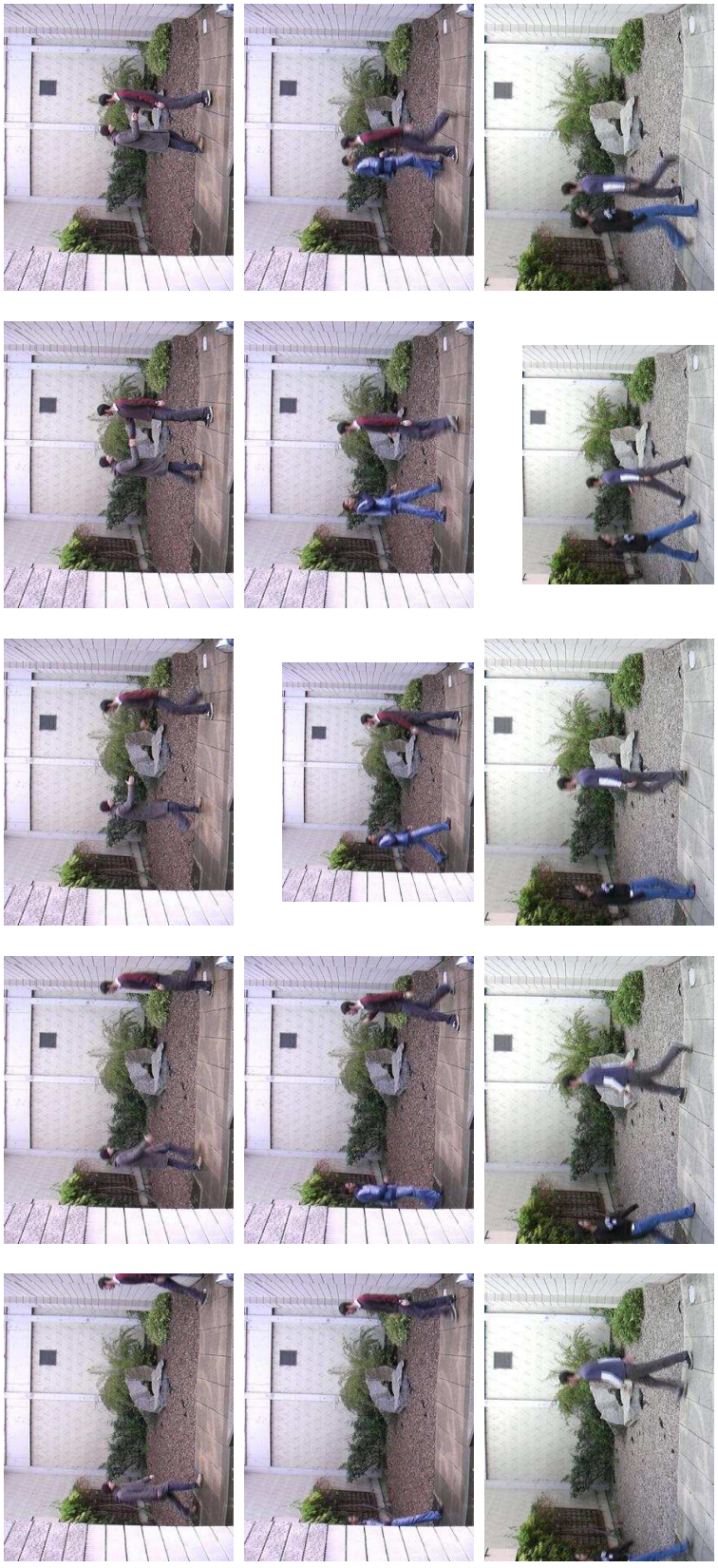


Figure 5.9: Examples of the 3 classes of interaction. The top row shows the shaking hands interaction, the next shows waving and the last shows just walking (or the non-interaction class).

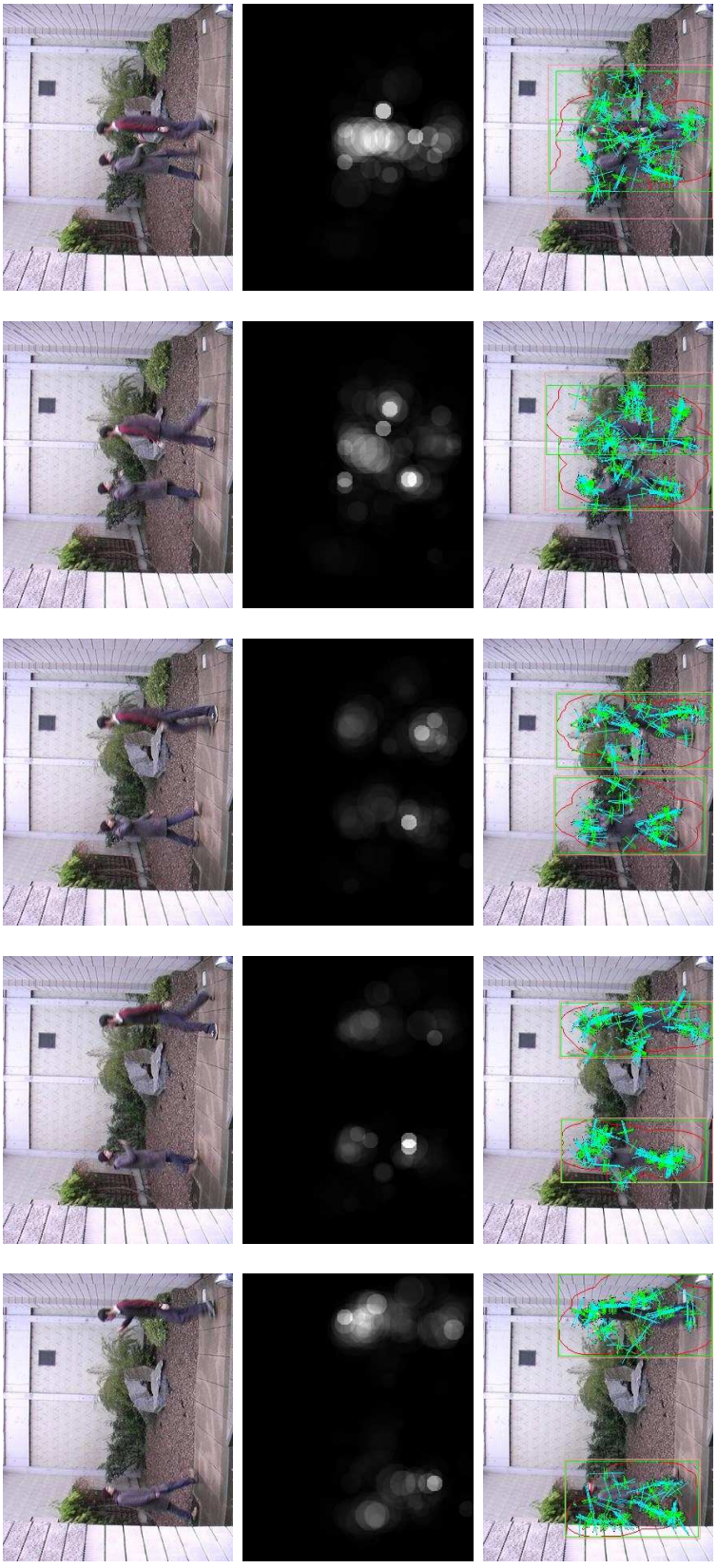


Figure 5.10: An example of the extracted features for the waving gesture. The first row shows the original sequence, the next shows the temporal saliency map and the final shows the grouping, estimated shape and pose information. The red lines indicate the boundaries of the spatially salient ellipses that were found from the corresponding temporal saliency map in the second row. The green boxes show the selection of the bounding boxes of both people based on the red line and also the most likely width of the bounding box for each person. The pink box shows the original hypothesis of where the bounding box was given only the salient ellipses extracted from the temporal saliency map. The cyan lines indicates orientation of the salient region of interest and the green lines crossing the cyan ones indicate the centroid of the salient region of interest that was used to calculated the orientation. The length of the two lines indicates the contours of the estimated elliptical volume.

Experiments were carried out using the leave-one-out method. Each class was evaluated by example so that the class which corresponded to the smallest distance between test and training data was considered to be the classified class. In total, there were 17 examples of non-interaction, 58 of shaking hands and 69 of waving. The classification results are shown in the table below:

		Test interactions		
		Shake hands	Wave	No interaction
Training interactions	Shake hands	3	5	2
	Wave	53	60	14
	No interaction	0	1	0

5.3.8 Discussion and Conclusion

The results are unsatisfactory. Under all circumstances, the model is always most likely to choose the waving gesture amongst all others. After some investigation, there are several different reasons for the performance of the results. Firstly, and perhaps the most significant problem is that the scale saliency method doesn't always perform well under considerable background clutter. By using colour, the number of possible intensity bins is increased considerably (for a reasonable representation of the colour space) which means that both low and high entropy values are less likely, leading to more flat entropy-scale characteristics. However, this also means that bright colours on a contrasting cluttered background are also more likely to be extracted compared to only grey-scale. On the otherhand, since the background is considerably cluttered, using just grey-scale is not suitably discriminative. Further investigations into perceptually uniform colour spaces may yield more promising results, though this will not remove the problem of sparsity given the higher dimensionality of the feature representation.

Another problem is that areas with high spatial homogeneity or that have an intensity histogram with low entropy tend to not be selected if the background region around it is too highly cluttered. Clutter in this sense is defined as any area which contains a relatively flat

intensity histogram. The problem is that as the scale increases, the entropy is required to reduce. However, if the area surrounding the region is extremely cluttered, then after the salient scale of the homogeneous region, the entropy will keep increasing. The scale saliency algorithm relies on the fact that the entropy must drop again at some point after the scale becomes greater than the salient region of interest. In some cases, given a large enough scale, the entropy does eventually drop. However, if the rest of the scales are also heavily cluttered, a salient scale may never be selected. This is illustrated in Figure 5.11 where a 100×100 circular region centred at the white cross indicated at the top of the figure, yields no salient region selection. In (a-d), the entropy-scale characteristics using 256×256 hue-saturation intensity bins, 16×16 hue-saturation intensity bins, 256 grey-level intensity bins and 16 grey-level intensity bins respectively, are shown. Even with the wide range of bins and calculating using just grey-level rather than colour yielded very similar results.

One possible solution is to increase the number of peaks in the entropy-scale characteristic. This requires us to remove the windowing function on the circular kernel described in Equation (5.10). This was originally used to increase the saliency value of regions which were better aligned with the centre of a homogeneous region. However, even if there are more peaks in the entropy-scale characteristic, this is no guarantee that the ranked global saliency will perform better. Without the windowing function, it is likely that the ordering of the salient regions will be less meaningful since selected regions of interest may not exactly overlap the true salient region, leading to many highly ranked values around the same region of interest. Another possible option may be to accumulate the entropy inversely so that the entropy-scale values are evaluated from histograms of the decreasing scale, starting with the outermost ring of the sampling kernel first.

Another problem which could be causing the results is to do with the sparsity of the co-occurrence matrix feature vectors. Despite fairly dense selection of spatio-temporally salient regions, the co-occurrence matrix, formed from Equation (5.23) will have sparse elements, particularly if the correlation between shapes is strong. This means that there was probably not enough non-singular elements to find a representative canonical space for the two sets of data. One possible solution would be to ignore this step and just compare the two feature

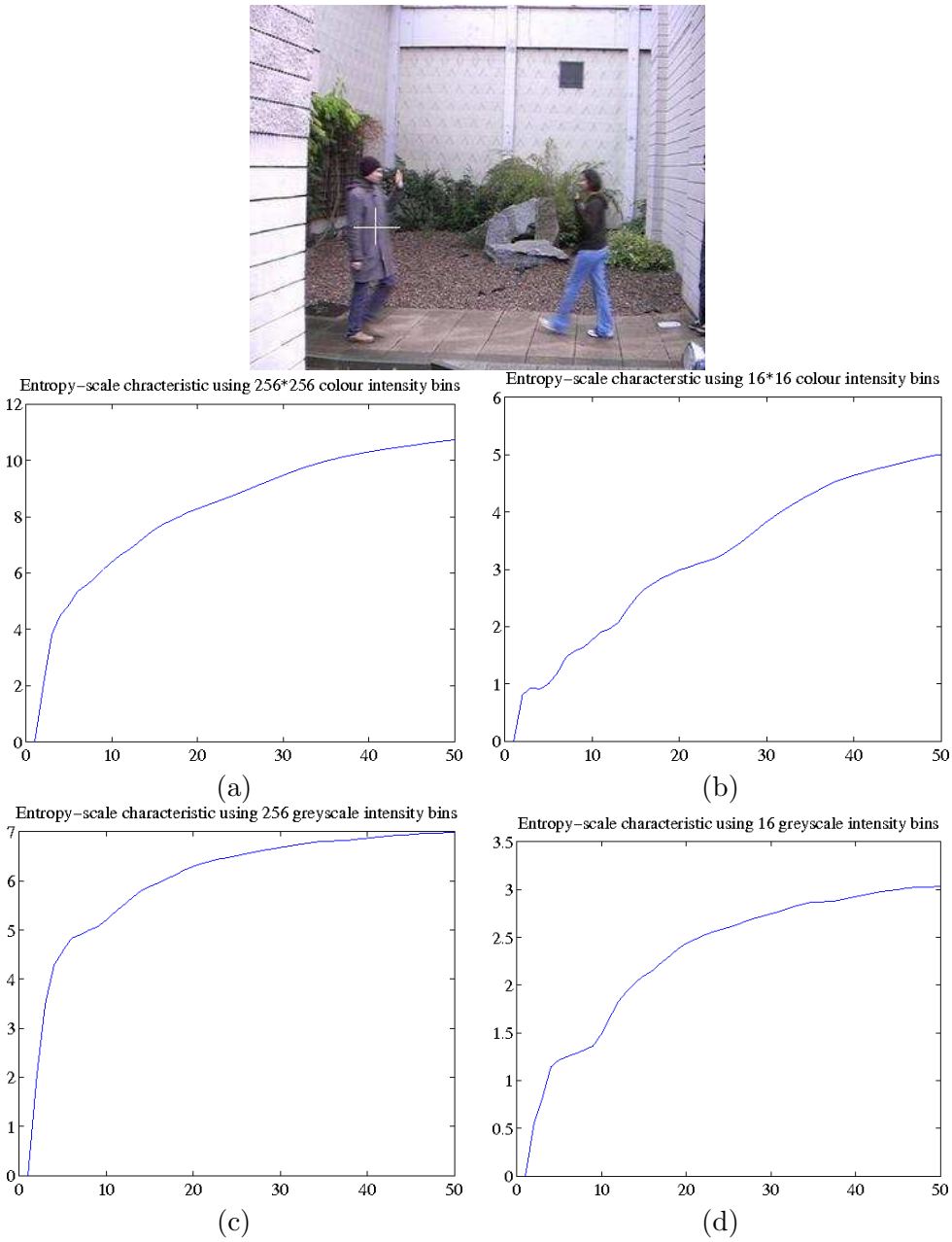


Figure 5.11: Graphs showing the limitation of the scale saliency algorithm in terms of the entropy-scale characteristics of the region centred with a cross in the top figure. (a) shows the entropy-scale characteristic using 256×256 hue-saturation intensity bins, (b) shows the same but for 16×16 hue-saturation intensity bins, (c) shows the same but for 256 grey scale bins, and (d) shows the characteristic for 16 bins. In each case, even though the region of interest is centred very clearly on a homogeneous region, it is not selected since the entropy-scale characteristic never peaks.

trajectories. However, this means that the recognition would be performed based on the assumption that person 1 and 2 are acting independently of each other. Another possibility would be to use a topic or aspect model approach such as probabilistic latent semantic analysis (Hofmann, 1999; Zalewski and Gong, 2004, 2005) to discover categories as latent interactions arising from the co-occurrence between feature vectors in the data itself rather than more specific predefined classes.

6 Conclusion and Futurework

The work in the previous chapters has provided an insight into the issues involved in interpreting the spatio-temporal dynamics of group and individual behaviour use the scene data alone. Whilst some interesting results have been gained, there are many open questions that remain. Saliency has been approached in the literature either as a spatial or spatio-temporal interest point detection process. However, for more complex activity, saliency detection is usually addressed as an abnormal behaviour or activity detection problem. However, this is perhaps a little too simplistic. Clearly there are many more subtle layers of saliency in the raw data that can be extracted which can provide a rich and meaningful framework for understanding the spatio-temporal dynamics of a particular scene, which is the focus of this thesis.

6.1 Temporal saliency as a measure of spatial plausibility

An important question that was highlighted in Chapter 5 is the relevance of spatial saliency, temporal saliency and spatio-temporal saliency. On the one hand, we expect that spatial saliency is more important in images and temporal saliency would be more significant for moving images. However, in this chapter, the need for spatio-temporal saliency became apparent for extracting salient people from the scene. Combining spatial and temporal saliency to produce a measure of the saliency of a spatio-temporal region is non-trivial and the weighting summation function that was used to combine both saliency terms in this chapter is perhaps rather arbitrary. One could apply more robust object recognition techniques to the problem at the first stage. But as yet, truly reliable detectors are not available and suffer from higher than tolerable false positive rates. Work has been carried out to combine both the

likelihood of the categorisation of an object as well as its likelihood of motion. However this relies heavily on the fact that salient objects in video will always be moving. In some cases, relatively stationary objects interacting with more or less temporally salient objects may be more significant but may have been filtered out at the early stages of feature extraction. So again, deciding on the range of multiple layers of context remains a challenging problem.

The assumption was made that on the whole, moving objects that were approximately elliptical could be treated as people in the scene. One could argue that this is quite a simplistic assumption since approximating a person as a moving vertically oriented ellipse ignores all current literature on person detection in crowded scenes. On the other hand, even the most successful person detection algorithms are prone to false positives that must be compensated for by using motion cues. The question arises as to whether this is a reasonable assumption to make given that not all people in a scene are necessarily moving. How do we know that the people that are just standing are not of equal interest? What makes moving people more salient?

The debate about whether spatial or temporal saliency is more important, or indeed how they can be combined in an accurate manner remains to be investigated. The work in Chapter 4 overcame this problem by binding locally, but not necessarily globally, salient homogeneous regions together to find potentially more salient interactions between them. This is quite an important issue in discovering salient patterns of activity since the process of binding objects or features may yield much more salient responses than the combined spatio-temporal saliency of the regions of interest. This is significant as it is not always considered when approaching saliency in video.

6.2 Saliency as a measure of rarity

Saliency has been quantified in this thesis as a measure of unpredictability or rarity of a particular feature distribution or statistical measure. However, this may not be the most reliable method of extracting saliency at all levels. As discussed in the previous section saliency is very much dependent on context and this is only defined in terms of whichever

feature space we choose to observe the representation of the data in. Ultimately, the feature space we decide to use is skewed in favour of a particular aspect of the data we are interested in quantifying. Therefore, in the case of the work investigated in this thesis, spatial or spatio-temporal saliency is blind to the feature binding process that is used at a later stage to find salient interactions. One could argue that this is an inevitable part of the feature extraction process but blindly removing less salient features at the early stages of spatio-temporal feature extraction may mean that salient activity formed from a combined set of less salient activities is missed. Therefore, we are still left with the need to manually set parameters to deal with this blind spot. Unfortunately, this then provides a much less clean and self-contained solution to the salient feature extraction problem and leads to some reliance on case-specific expert knowledge.

Issues that remain to be solved in the area of temporal saliency addresses scoping. Is something which is salient within the last five minutes still salient in within the last day or month? Furthermore, if something is temporally salient in the local spatial neighbourhood, is it still temporally salient in the global spatial context? There are clearly many levels of context at which an object comes less or more relevant to the interpretation of a scene. So far, discussions have centred around the idea of temporal saliency in terms of salient motion patterns. However, this is clearly not enough to understand more complex cause and effect phenomenon from natural scenes. What we must turn towards is the idea of how to combine these salient features in a descriptive but discriminative manner in order to understand more realistic natural human behaviour.

6.3 Tackling challenging view angles

Is it necessary? Perhaps in many cases the easiest and cheapest solution to major occlusion problems is to move the camera. However, until automated systems become more dominant in the surveillance domain, it is unrealistic to expect camera placement to be informed by the limitations of automation techniques except in very specific computer vision problems such as people counting. In any case, for more crowded scenes, interpreting activity still requires

a camera to capture a person from an approximate side-angle which will ultimately lead to problems of self occlusion as well as occlusion from topology or objects within the scene.

The idea presented in Chapter 4 of using co-occurrence to find interactions bypasses the need to use tracking. However, using saliency as the co-occurrence term threw away considerable information to do with the local spatio-temporal structures of the regions of interest. Therefore, in Chapter 5 a more specific application was given with the idea of finding interactions by looking strictly at the interaction of people. Here, saliency was found by events generated from the body behaviour of detected spatio-temporally salient objects within the scene. The idea of using events to find interactions is an effective method of taking advantage of cues wherever we can find them. In this way, instead of looking for trained models of particular activities, we can search for locally temporally salient events, which is only dependent on the context of the local temporal scene data. The disadvantage of the method described in Chapter 5 as opposed to 4 is that there is a heavy reliance on object tracking for feature alignment and also trajectory event detection. Unfortunately, this is an inevitable part of the event detection process. However, in most cases, as long as the scene is not too crowded, some object tracking is possible and sometimes only a few frames were needed before enough evidence was gathered to find events in the local temporal neighbourhood. Also, at some point, there must be a departure from non-tracking based methods since otherwise, the extracted features become salient in terms of a receptive field, which is spatially translation variant. Clearly to perform realistic group behaviour detection we must remove dependence on scene topology based feature representation methods such as those shown in (Chomat et al., 2000).

A significant weakness of the work in this thesis is that effective scene interpretation relies on a relatively non-occluded viewing angle. However, in real natural surveillance scenes, occlusion is still inherent. I argue strongly that many of the cameras that are used for surveillance using human operators are not suitable for automated surveillance. A well chosen camera position can greatly increase the performance of any automated surveillance system. Clearly if we want to count the number of people walking into a shopping mall on a busy weekend afternoon, a camera mounted with a top down view would be much more effective for

head counting. Whilst monitoring behaviour where gestures or body actions are important would benefit more from a side view.

Despite the disappointing results in the previous chapter, there are many interesting questions raised about how temporally correlated data can be represented in terms of co-occurrences. In particular, despite the overriding theme in this thesis that it is important to extract salient features before performing more semantically meaningful binding processes, at each stage, we were only able to extract what was salient because a model of the less salient features were present and could provide some discriminance. This was demonstrated in both Chapters 3 and 4 where a considerable amount of background information needed to be selected in order to find the more salient information from the scenes. Furthermore, in the previous chapter, there was a problem with sparsity in the co-occurrence matrices, which meant that using linear methods such as CCA was not particularly effective. If the co-occurrence matrices were populated with co-occurrences from the orientations of shapes in the background, it might have yielded more reliable measure of the co-occurrence between the co-occurrence vectors of the two people in the sequences.

One might argue that more robust approach might be the work of (Park and Aggarwal, 2003) which uses a hierarchical Bayesian model of the shape and head-pose information from the human body in order to classify different types of actions. However, the major drawback of such a method is that the it is highly supervised through the design of its complex hierarchical Bayesian network of each human body. Much of the plausibility of different limbs being present in the scene rely on these causal models, which is unlikely to generalise to real outdoor environments where coats may obscure limbs and background clutter is very likely. Research in this area has tended to be forced into one of two areas where the scene and action data is highly simplified, often relying on silhouette-based approaches, or the classification of activities are addressed instead. Making a compromise in the direction of the first approach is more likely to lead to methods that will never work in a practical situation since the input features are too simple and the corresponding classification methods rely heavily on the nature of the feature representation. The latter is perhaps a more realistic method but means that detecting subtle behaviour becomes more difficult since classifying

activities implies that the feature representation does not necessarily need to have a possible higher semantic meaning. Overall, unless the right feature and binding if found to deal with real situations, it is perhaps only an exercise in interesting research ideas rather than real solutions to a very significant problem in video behaviour understanding.

6.4 Further work

The results and ideas presented in this thesis provides only the beginning of many interesting avenues for further study.

Representing direction of motion This was touched on briefly when the idea of using a cylindrical kernel as opposed to an elliptical kernel (extending backwards in time) was suggested in order to remove any bias to motion in any particular direction. This means that with a carefully chosen ellipse ratio, it would be possible to measure dominant directions of motion of a temporally salient region. This would provide a neat addition to the temporal saliency algorithm and lead to a richer descriptor of the selected area. It could be used to replace the local spatio-temporal binding using triples described in Chapter 4 in Section 4.1. The original spatio-temporal binding that was presented acts as a method of accumulating likely correspondences between frames of a salient region of interest which we assume to be a single object. Using a more specific motion descriptor would lead to less ambiguity between correspondences.

Detecting interacting people from an aerial view None of the work presented in here used aerial views of humans in order to classify activity since limb movements are difficult to distinguish under such viewing conditions. However, it would be interesting to apply Ullman and Shashua (1988)'s structural saliency to aerial views of static formations of interacting people in order to identify salient group formations. Drawing from Kendon (1990)'s findings that the spatial formation of people varies depending on their level of engagement with each other, it would be interesting to see if statistical methods would reach the same conclusions.

Reverse histogram sampling for more robust salient region detection Given the problems with using Kadir and Brady's scale saliency algorithm for extracted salient homogeneous spatial regions of interest in cluttered scenes, it is necessary to re-address the intensity histograms are sampled over increasing scale. As explained in the previous chapter, if the entropy-scale characteristic were calculated for a histograms of decreasing scale, so that the model would be accumulated from the outer edges of the circular kernel inwards, it might be possible to find method which could handle salient homogeneous objects with very similar colour distributions to the background.

Bibliography

D. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, 2(2):284–299, 1985.

A. Ali and J. K. Aggarwal. Segmentation and recognition of continuous human activity. In *IEEE Workshop on Detection and Recognition of Events in Video*, pages 28–, 2001.

A. Bar-Hillel, T. Hertz, and D. Weinshall. Efficient learning of relational object class models. In *International Conference on Computer Vision*, pages 1762–1769, 2005.

J. Benjafield. *The developmental point of view. A history of psychology.* Simon and Schuster Company, 1996.

J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *Proceedings of the Second European Conference on Computer Vision*, pages 237–252, 1992.

T. Binford. Inferring surfaces from images. *Journal of Artificial Intelligence*, 17(1-3):205–244, August 1981.

C. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, England, 1995.

M. J. Black and A. D. Jepson. Recognizing temporal trajectories using the condensation algorithm. In *FG '98: Proceedings of the 3rd. International Conference on Face & Gesture Recognition*, page 16, Washington, DC, USA, 1998. IEEE Computer Society. ISBN 0-8186-8344-9.

A. Bobick and J. Davis. The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence*, 23(3):257–267, March 2001.

O. Boiman and M. Irani. Detecting irregularities in images and in video. In *International Conference on Computer Vision and Pattern Recognition*, 2005.

O. Boiman and M. Irani. Similarity by composition. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.

M. Burl, T. Leung, and P. Perona. Face localization via shape statistics. In *Intl. Workshop on Automatic Face and Gesture Recognition*, June 1995.

P. Chandon, J. W. Hutchinson, E. T. Bradlow, and S. Young. *Measuring the Value of Point-of-Purchase Marketing with Commercial Eye-Tracking Data*. Visual Marketing:From Attention to Action. Lawrence Erlbaum Associates, 2007.

O. Chomat and J. Crowley. Recognizing motion using local appearance. In *International Symposium on Intelligent Robotic Systems*, 1998.

O. Chomat, J. Martin, and J. Crowley. A probabilistic sensor for the perception and recognition of activities. In *European Conference on Computer Vision*, pages 487–503, 2000.

L. Davis, S. Johns, and J. Aggarwal. Texture analysis using generalized co-occurrence matrices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(3):251–259, July 1979.

L. Davis, M. Clearman, and J. Aggarwal. An empirical evaluation of generalized cooccurrence matrices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3(2):214–221, March 1981.

A. Davison. Real-time simultaneous localisation and mapping with a single camera. In *International Conference on Computer Vision*, 2003.

A. Davison. Active search for real-time vision. In *International Conference on Computer Vision*, 2005.

H. Dee and D. Hogg. Detecting inexplicable behaviour. In *British Machine Vision Conference*, 2004.

K. Derpanis and J. Gryn. Three-dimensional nth derivative of gaussian separable steerable filters. In *International Conference Image Processing*, volume 3, pages 553–556, 2005.

I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Knowledge Discovery and Data Mining*, pages 269–274, 2001.

T. Doll and R. Home. Guidelines for developing and validating models of visual serach and target acquisition. *Optical Engineering*, 40:1776–1783, Sept. 2001.

C. Donald. Assessing the human vigilance capacity of control room operators. In *In International Conference on Human Interfaces in Control Rooms, Cockpits and Command Centres*, pages 7–11, 1999.

G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto. Dynamic textures. *International Journal of Computer Vision*, 51(2), 2003.

A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *International Conference on Computer Vision*, 2003.

I. Elfadel and R. Picard. Gibbs random fields, cooccurrences, and texture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(1):24–37, 1994.

H.-L. Eng, K.-A. Toh, A. H. Kam, J. Wang, and W.-Y. Yau. An automatic drowning detection surveillance system for challenging outdoor pool environments. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 532, 2003.

L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594, 2006.

W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.

A. Galata, N. Johnson, and D. Hogg. Learning variable length markov models of behaviour. *Computer Vision and Image Understanding, CVIU*, 81(3):398–413, 2001.

A. Galata, A. Cohn, D. Magee, and D. Hogg. Modelling interaction using learnt qualitative spatio-temporal relations and variable length markov models. In *European Conference on Artificial Intelligence*, 2002.

M. A. Giese and T. Poggio. Neural mechanisms for the recognition of biological movements and action. *Nature Reviews Neuroscience*, 4:179–192, 2003.

J. S. . S. Gong. Vigour: A system for tracking and recognition of multiple people and their activities. In *International Conference on Pattern Recognition*, volume 1, Computer Vision and Image Analysis, September 2000.

S. Gong and M. Brady. Parallel computation of optic flow. In *ECCV 90: Proceedings of the first european conference on Computer vision*, pages 124–133, New York, NY, USA, 1990. Springer-Verlag New York, Inc.

S. Gong and T. Xiang. Recognition of group activities using dynamic probabilistic networks. *International Conference of Computer Vision*, 02:742, 2003.

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

T. X. . S. Gong. Activity-based scene representation and segmentation of cctv surveillance videos. In *British Machine Vision Conference*, September 2004.

C. Gotleib and H. Kreyszig. Texture descriptors based on co-occurrence matrices. *Computer Vision, Graphics and Image Processing*, 51(1):70–86, 1990.

G. H. Granlund. In search of a general picture processing operator. *Computer Graphics and Image Processing (CGIP)*, 2:155–173, 1978.

A. Graves and S. Gong. Spotting scene change for indexing surveillance video. In *British Machine Vision Conference*, 2003a.

A. Graves and S. Gong. Spotting scene change for indexing surveillance video. In *British Machine Vision Conference*, pages 469–478, Norwich, England, September 2003b.

A. Hakeem and M. Shah. Ontology and taxonomy collaborated framework for meeting classification. *International Conference on Pattern Recognition*, 04:219–222, 2004.

R. Hamid, A. Johnson, S. Batta, A. Bobick, C. Isbell, and G. Coleman. Detection and explanation of anomalous activities: Representing activities as bags of event n-grams. In *International Conference on Computer Vision and Pattern Recognition*, 2005.

R. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67 (5):786–803, May 1979.

C. Harris and M. Stephens. Combined corner and edge detector. In *Proceedings of the Fourth Alvey Vision Conference*, pages 147–151, 1988.

E. Hildreth. Computing the velocity field along contours. In *SIGGraph/SIGArt Interdisciplinary Workshop on Motion:Representation and Perception*, pages 26–32, 1983.

T. Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, pages 50–57, Berkeley, California, August 1999.

D. C. Hogg, N. Johnson, R. Morris, D. Buesching, and A. Galata. Visual models of interaction. In *2nd International Workshop on Cooperative Distributed Vision*, 1998.

B. Horn. The binford-horn line finder. In *MIT AI*, 1971.

B. Horn and B. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, August 1981.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 8:321–377, 1936.

D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *J Physiol.*, 148(3):574–591, 1959.

D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol.*, 160(1):106–154.2, 1962.

M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.

L. Itti and C. Koch. Computational modelling of visual attention. *National Review Neuroscience*, 2(3):194–203, March 2001. ISSN 1471-003X.

L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Journal of Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

J. H. . H. B. J. Sherrah. Interpretation of group behaviour in visually mediated interaction. In *International Conference on Pattern Recognition*, volume 1, Computer Vision and Image Analysis, pages 266–269, September 2000.

G. Johansson. Visual motion perception. *Scientific American*, 232:76–88, 1975.

N. Johnson and D. Hogg. Representation and synthesis of behaviour using gaussian mixtures. *Image and Vision Computing*, 20:889–894, 2000.

N. Johnson and D. Hogg. Learning the distribution of object trajectories for event recognition. In *British Machine Vision Conference*, 1995.

N. Johnson, A. Galata, and D. Hogg. The acquisition and use of interaction behaviour models. In *International Conference on Computer Vision and Pattern Recognition*, 1998.

T. Kadir. *Scale Saliency and Scene Description.* Phd dissertation, University of Oxford, 2002.

T. Kadir and M. Brady. Scale, saliency and image description. *International Journal of Computer Vision*, 45(2):83–105, November 2001.

T. Kadir, D. Boukerroui, and J. Brady. An analysis of the scale saliency algorithm. *Technical Report*, OUEL No. :2264/03, 2003.

T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. In *European Computer Conference on Computer Vision*, pages Vol I: 228–241, 2004.

A. Kendon. *Conducting Interaction: Patterns of Behaviour in Focused Encounters.* Cambridge University Press, Cambridge, UK, 1990.

Z. Khan, T. Balch, and F. Dellaert. An mcmc-based particle filter for tracking mutiple interacting targets. *GIT Technical Report*, GIT-GVU-03-35, Oct. 2003.

C. Koch and S. Ullman. Selecting one among the many: A simple network implementing shifts in selective visual attention. Technical report, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, and Center for Biological Information Processing, Whitaker College, Cambridge, MA, USA, 1984.

J. B. Kruskal and M. Liberman. *The symmetric time-warping problem: From continuous to discrete.* Time Warps, String Edits, And Macromolecules:The Theory and Practice of Sequence Comparison. CSLI Publications, 1999.

I. Laptev and T. Lindeberg. Space-time interest points. In *International Conference on Computer Vision*, 2003.

T. Lindeberg. Feature detection with automatic scale selection. *International Journal on Computer Vision*, 30(2), 1998a.

T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):77–116, 1998b.

D. Lowe. Local feature view clustering for 3d object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 682–688, December 2001.

D. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, pages 1150–1157, September 1999.

D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference Artificial Intelligence*, pages 674–679, 1981.

Y.-F. Ma and H.-J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 374–381, New York, NY, USA, 2003. ACM Press. ISBN 1-58113-722-2.

M. McCahill and C. Norris. Cctv in london. Technical Report Working Paper No.6, June 2002.

S. McKenna and H. Nait-Charif. Summarising contextual activity and detecting unusual inactivity in a supportive home environment. *Pattern Analysis and Applications*, 7(4): 386–401, December 2004.

S. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. Tracking groups of people. *Computer Vision and Image Understanding*, 80:42–56, 2000.

S. J. McKenna and S. Gong. Gesture recognition for visually mediated interation using probabilistic event trajectories. In *British Machine Vision Conference*, 1998.

A. Mittal and N. Paragios. Motion-based background subtraction using adaptive kernel density estimation. In *In IEEE Conference on Computer Vision and Pattern Recognition*, July 2004.

C. J. Needham and R. D. Boyle. Performance evaluation metrics and statistics for positional tracker evaluation. In *Computer Vision Systems: Third International Conference, ICVS 2003, Graz, Austria*, pages 278–289. Springer Verlag, April 2003.

U. Neisser. Visual search. *Scientific American*, 210(6):94–102, 1964.

J. Ng and S. Gong. Learning intrinsic video content using levenshtein distance in graph partitioning. *Lecture Notes in Computer Science*, 23(53):670–684, 2002.

J. Ng and S. Gong. Learning pixel-wise signal energy for understanding semantics. *Image and Vision Computing*, 21(12-13):1183–1189, December 2003.

S. Oh, S. Russell, and S. Sastry. Markov chain monte carlo data association for general multiple-target tracking problems. In *IEEE Conference on Decision and Control*, 2004.

A. Oikonomopoulos, I. Patras, and M. Pantic. Spatiotemporal saliency for human action recognition. In *IEEE International Conference on Multimedia and Expo 2005*, pages 430–433, July 2005. ISBN 0-7803-9332-5.

A. Oikonomopoulos, I. Patras, and M. Pantic. Kernel-based recognition of human actions using spatiotemporal salient points. In *IEEE International Conference on Computer Vision and Pattern Recognition 2006*, volume 3, June 2006a. ISBN 0-7695-2597-0.

A. Oikonomopoulos, I. Patras, and M. Pantic. Human action recognition with spatiotemporal salient points. *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, 36(3):710–719, June 2006b. ISSN 1083-4419.

N. Oliver, B. Rosario, and A. Pentland. A bayesian computer vision system for modeling human interactions. *Journal of Pattern Analysis and Machine Intelligence*, 22(8), 2000.

E. Ong and S. Gong. The dynamics of linear combinations: Tracking 3d skeletons of human subjects. *Image, Vision and Computing*, 20(5-6):397–414, March 2002.

S. Park and J. K. Aggarwal. Recognition of two-person interactions using a hierarchical bayesian network. In *IWVS '03: First ACM SIGMM international workshop on Video surveillance*, pages 65–76, 2003.

A. Pikaz and I. Disntein. Using simple decomposition for smoothing and feature point detection of noisy digital curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(8):808–813, 1994.

C. Rao, A. Yilmaz, and M. Shah. View-invariant representation and recognition of actions. *Internation Journal of Computer Vision*, 50(2):203–226, 2002.

E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In *European Conference on Computer Vision*, May 2006.

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

C. Salzman, C. Murasugi, K. Britten, and W. Newsome. Microstimulation in visual area mt: effects on direction discrimination performance. *Journal of Neuroscience*, 12:2331–2355, 1992.

B. Schiele and J. Crowley. Object recognition using multidimensional receptive field histograms. In *European Conference on Computer Vision*, pages 610–619, 1996.

C. Schmid, R. Mohr, and C. Bauckhage. Comparing and evaluating interest points. In *Proceedings of the International Conference on Computer Vision*, pages 230–235, 1998.

M. Seki, T. Wada, H. Fujiwara, and K. Sumi. Background subtraction based on cooccurrence of image variations. *International Conference on Computer Vision and Pattern Recognition*, 02:65, 2003.

L. Shafarenko, M. Petrou, and J. Kittler. Automatic watershed segmentation of randomly textured color images. *IEEE Transactions on Image Processing*, 6(11):1530–1544, 1997.

C. Shan, S. Gong, and P. W. McOwan. Capturing correlations among facial parts for facial expression analysis. In *Proc. British Machine Vision Conference (BMVC'07)*, 2007.

J. Sherrah and S. Gong. Tracking discontinuous motion using bayesian inference. In *European Conference on Computer Vision*, pages 150–166, 2000.

J. Sherrah and S. Gong. Exploiting context in gesture recognition. In P. Bouquet, L. Serafini, P. Brézillon, M. Benerecetti, and F. Castellani, editors, *Modeling and Using Contexts: Proceedings of the Second International and Interdisciplinary Conference, CONTEXT'99*, pages 515–518. Springer-Verlag, Berlin, 1999.

J. Shi and C. Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94)*, Seattle, 1994.

Y. Song, L. Goncalves, and P. Perona. Unsupervised learning of human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):814–827, 2003.

A. Spinei, D. Pellerin, and J. Herault. Spatiotemporal energy-based method for velocity estimation. *Signal Processing*, 65:347–362, 1998.

C. Stauffer and E. Grimson. Learning patterns of activity using real-time tracking. *Journal of Pattern Analysis and Machine Intelligence*, 22(8), 2000.

C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, 1999.

K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *International Conference on Computer Vision*, pages 255–261, 1999.

A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

T. Troscianko, A. Holmes, J. Stillman, M. Mirmehdi, and D. Wright. Will they have a fight? the predictability of natural behaviour viewed through cctv cameras. In *European Conference on Visual Perception 2001, Perception Vol 30 Supplement*, pages 72–72. Pion Ltd, August 2001.

S. Ullman and A. Shashua. Structural saliency: The detection of globally salient structures using a locally connected network. Technical Report AIM-1061, Weizmann Institute, Cambridge, MA, USA, 1988.

M. Usher and N. Donnelly. Visual synchrony affects binding and segmentation in perception. *Nature*, 394(6689):179–182, July 1998.

W. R. Uttal. *An autocorrelation theory of form detection*. L. Erlbaum Associates ; New York : distributed by Halsted Press, Hillsdale, N.J., 1975.

M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *European Conference on Computer Vision*, pages 18–32, 2000.

M. Wertheimer. *Classics in Psychology-Experimental Studies on the Seeing of Motion*. US/Mountain, 1961.

E. Williams. Visual search: A novel psychophysics for preattentive vision. Master's thesis, Research Science Institute, August 1999.

L. Wixson. Detecting salient motion by accumulating directionally-consistent flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 2000.

C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder:real-time tracing of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7): 780–785, July 1997.

T. Xiang and S. Gong. Online video behaviour abnormality detection using reliability measure. In *British Machine Vision Conference*, 2005a.

T. Xiang and S. Gong. Video behaviour profiling and abnormality detection without manual labelling. In *International Conference on Computer Vision*, volume 2, pages 1238–1245, 2005b.

T. Xiang and S. Gong. Beyond tracking: Modelling activity and understanding behaviour. *International Journal on Computer Vision*, 67(1):21–51, 2006.

T. Xiang, S. Gong, and D. Parkinson. Autonomous visual events detection and classification without explicit object-centred segmentation and tracking. In *British Machine Vision Conference*, 2002.

L. Zalewski and S. Gong. 2d statistical models of facial expressions for realistic 3d avatar animation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 217–222, 2005.

L. Zalewski and S. Gong. Synthesis and recognition of facial expressions in virtual 3d views. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 493–498, 2004.

D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, and G. Lathoud. Multimodal group action clustering in meetings. In *VSSN '04: Proceedings of the ACM 2nd international workshop on Video surveillance & sensor networks*, pages 54–62, 2004. ISBN 1-58113-934-9.

T. Zhao and R. Nevatia. Tracking multiple humans in crowded environment. In *Computer Vision and Pattern Recognition*, 2003.

H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *International Conference on Computer Vision and Pattern Recognition*, 2004.

7 Appendices

7.1 Initial study of the problem

This chapter serves to highlight in more detail some of the problems relating to understanding activity in video. As an initial experiment, we shall deal with the problem of classifying human features in order to better understand the problem domain. For the purposes of these experiments, many conditions have been controlled in order to try and maximise the accuracy of the recognition process.

- Problems of background clutter have been removed by the use of a blue screen in the background of all the sequences.
- The lighting has been kept as constant as possible in both direction and intensity.
- All the subjects sit while their arms perform the gestures so that the rest of their body remains relatively stationary.
- Only adults are used for the data so they are approximately of the same size.
- All gestures are temporally segmented manually.

Figure 7.1 shows a key frame of each gesture that was used in the experiments presented in this chapter.

This chapter is primarily divided into two parts. The first part uses a simple feature extraction method to find a representation of the overall motion in each frame of the gesture sequences. From this representation, a set of feature trajectories are warped locally for matching and classification. The Condensation algorithm, originally proposed by Isard and

Figure 7.1: The 6 gestures that were used for recognition. Clockwise from top left;come, go, point left, point right, a high wave and a low wave.

Blake (1998), will be used to perform the classifying where the implementation was based on the work of Black and Jepson (1998). In the second part, the feature extraction is modified to represent the orientation and spatial location of the motion and rather than using trajectories, the features are accumulated over time into a single oriented motion template using the multi-dimensional receptive field histograms suggested by Chomat and Crowley (1998).

7.1.1 Feature Extraction and Tracking

In order to investigate the matching properties of the Condensation algorithm, the simplest form of feature extraction, through motion moments, was used (McKenna and Gong, 1998). Motion moments are obtained using two-frame temporal differencing. This involves taking the difference between consecutive frames and creating a binary image $\mathbf{B}_t$. 0s and 1s are determined by the difference between pixel values in consecutive frames, being above or below a predefined threshold. This technique assumes that the subject remains relatively motionless so that the major change in movement is due to the gesture-making. Once a binary image is created, it is used as a mask to calculate the total motion area A_t, displacement of the centroid coordinates u_t and v_t, and elongation E_t (derived from the second order moments).

These features are calculated using the following equations, where x and y represent the

coordinate values of the 2-dimensional image at each time frame:

1. Motion area:

$$A_t = \sum_{x,y} \mathbf{B}_t[x, y] \tag{7.1}$$

2. Centroid coordinates of the motion area:

$$\overline{x}_t = \frac{1}{A_t} \sum_{x,y} x\mathbf{B}_t[x, y] \tag{7.2}$$

$$\overline{y}_t = \frac{1}{A_t} \sum_{x,y} y\mathbf{B}_t[x, y] \tag{7.3}$$

3. Displacement of centroid coordinates of the area of motion:

$$u_t = \overline{x}_t - \overline{x}_{t-1} \tag{7.4}$$

$$v_t = \overline{y}_t - \overline{y}_{t-1} \tag{7.5}$$

4. Elongation of the area of motion:

$$E_t = \frac{\chi_{max}}{\chi_{min}} \tag{7.6}$$

where, $\chi^2 = \frac{1}{2}(\mathrm{a} + \mathrm{c}) + \frac{1}{2}(\mathrm{a} - \mathrm{c})cos2\theta + \frac{1}{2}\mathrm{b}sin2\theta$

$\mathrm{a} = \sum_{x,y}(x - \overline{x})^2\mathbf{B}_t[x, y]$,

$\mathrm{b} = \sum_{x,y}(x - \overline{x})(y - \overline{y})\mathbf{B}_t[x, y]$,

$\mathrm{c} = \sum_{x,y}(y - \overline{y})^2\mathbf{B}_t[x, y]$,

$sin2\theta = \pm\frac{\mathrm{b}}{\sqrt{\mathrm{b}^2+(a-\mathrm{c})^2}}$,

$cos2\theta = \pm\frac{\mathrm{a}-\mathrm{c}}{\sqrt{\mathrm{b}^2+(\mathrm{a}-\mathrm{c})^2}}$

The maximum and minimum levels of χ are found by altering the signs of the cos2θ and sin2θ expressions. Hence the resultant feature set is: (A_t, u_t, v_t, E_t)

From the description of this feature extraction technique, we can see that tracking is also incorporated into this, with the use of the displacement features u and v, which show the general relative movement of the most significant parts of the body, during the performance of a gesture. Using these features, a representation of the temporal evolution of each class is possible by concatenating this feature set at each time frame into a multi-dimensional feature vector trajectory.

7.1.2 The Condensation algorithm

The advantage of the Condensation algorithm is that it is able to propagate the probability densities of many states in the search space at every iteration. Therefore, it is particularly useful for multiple tracking problems as well as gesture recognition. For simplicity, the technique used here takes the overall area of motion at each frame to be one point in the motion trajectory vector. However, it already becomes apparent that some classes that require motion from 2 spatially separated hand motions will lead to inaccurate representations of the gesture trajectory. At this stage, we could try to identify multiple areas of motion but this would lead to an non-trivial increase in the complexity of this problem and is out of the scope of this thesis.

Before the implementation of the Condensation algorithm is described, it is important to clarify some of the terminology that will be used. The application of this algorithm has been based on an implementation for the recognition of temporal trajectories on a 2-dimensional surface (Black and Jepson, 1998). Thus, much of the terminology has been kept from this paper. In this context, the word 'trajectory' has been used to mean one of the dimensions from the feature set, which varies over the time of the gesture sequence. For example, the variation of the motion area over time for one particular gesture is considered to be a trajectory in feature space.

In each class, the same gesture will never be identical. In order to overcome this, it must be possible to distort the feature vector trajectories to enhance discriminance between the class models and the test gesture sequence. This allows generalisation of a particular gesture by distorting each feature trajectory by 3 different parameters; α, ρ, and ϕ which represent

amplitude, rate and phase adjustments within a local time window. Each state in the search space contains values for these 3 variables, as well as a variable μ to represent the model or gesture. μ generally represents a different number, for each gesture. A state at time t is defined as $\mathbf{s}_t = (\mu, \phi, \alpha, \rho)$.

Essentially, the Condensation algorithm consists of 4 basic steps; initialisation, selection, prediction and updating, as shown in Figure 7.2. A brief description of the Condensation algorithm can be found in Black and Jepson (1998). A more detailed and extended description of the algorithm is provided as follows. Figure 7.3 illustrates the basic flow diagram of the

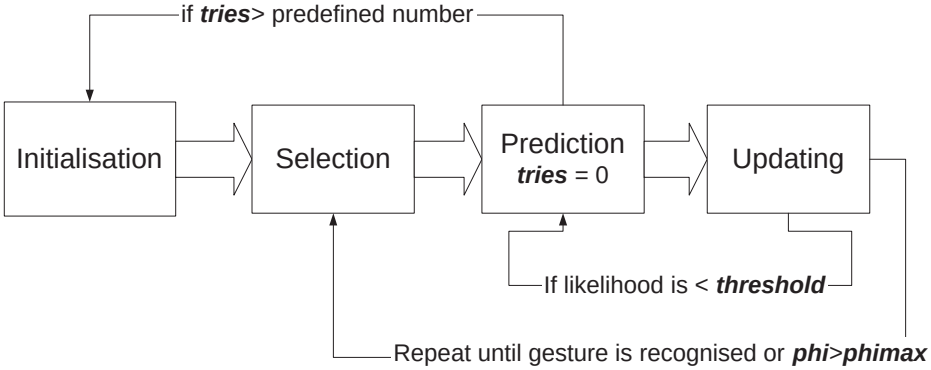


Figure 7.2: High level block diagram of the Condensation algorithm.

Condensation algorithm, used for gesture recognition. Each trajectory described in this figure represents the variation of one particular feature over time. The training data was treated as a vector $\mathbf{m}$ of N values, where $\mathbf{m}^{(\mu)} = (m_1^{(\mu)}, ..., m_N^{(\mu)})$.

The search space was firstly initialised by choosing S sample states (typically of the order of 1000). This produced a set $\{\mathbf{s}_t^{(n)}, n = 1, ..., S\}$ samples. The purpose of the algorithm is to find the most likely state $\mathbf{s}_t$ that creates the best match for the input or observation data, $Z_t = (\mathbf{z}_t, \mathbf{z}_{t-1}, ...)$. The observation vector for a particular trajectory i (or each variable of the feature set) is $Z_{t,i} = (z_{t,i}, z_{(t-1),i}, z_{(t-2),i}, ...)$. To find likelihoods for each state, DTW, according to the state parameters, must be performed on the model data. This is calculated as the probability of the observation $\mathbf{z}_t$ given the state $\mathbf{s}_t$ and is given by:

$$p(\mathbf{z}_t \mid \mathbf{s}_t) = \prod_{i=0}^{N} p(Z_{t,i} \mid \mathbf{s}_t), \tag{7.7}$$

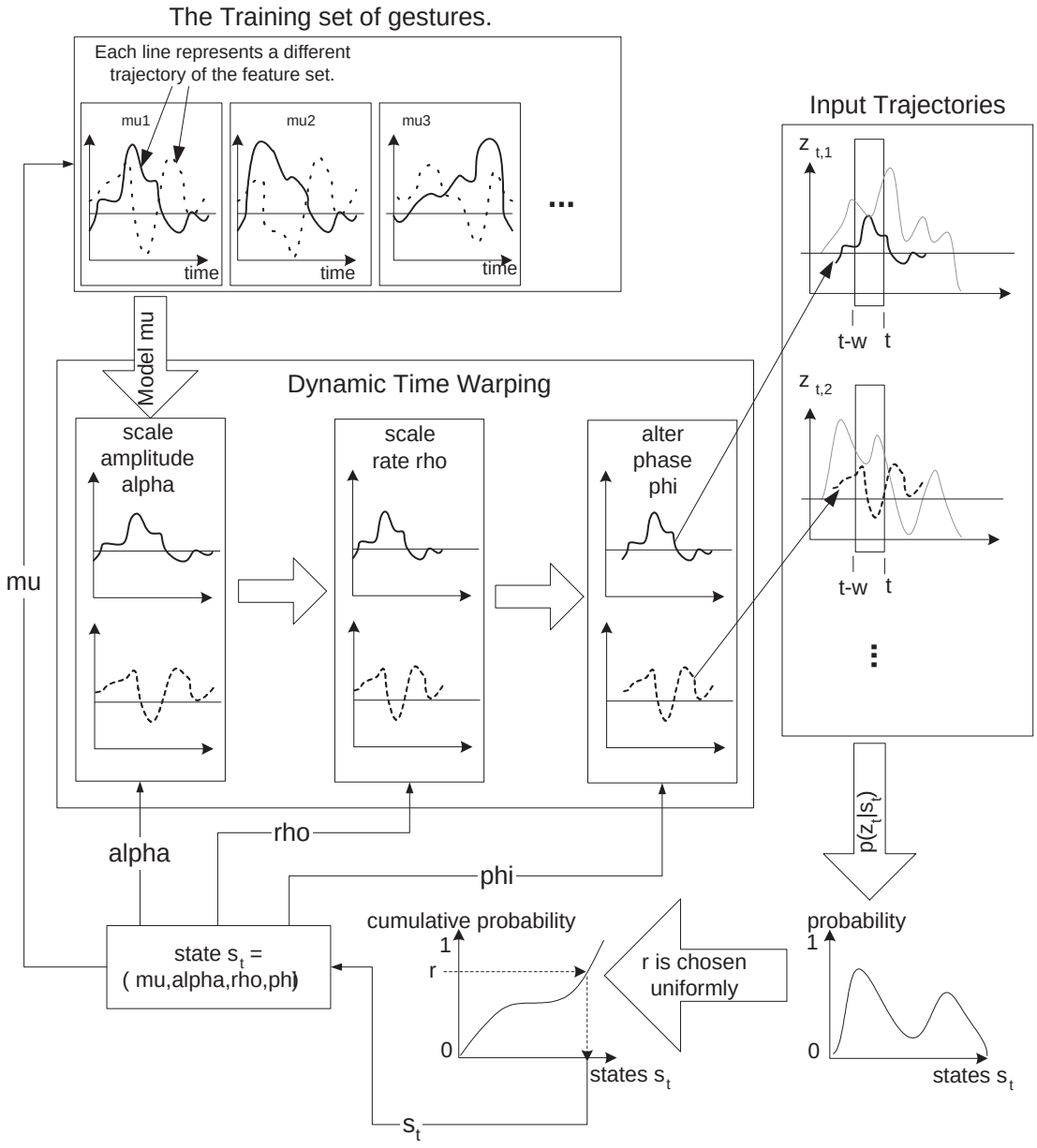


Figure 7.3: Flow Diagram of the matching and selection processes of the Condensation algorithm.

where

$$p(Z_{t,i}|s_t) = \frac{1}{\sqrt{2\pi\sigma_i}} exp \frac{-\sum_{j=1}^{\omega-1} \left(z_{(t-j),i} - \alpha m^{(\mu)}_{(\phi-\rho j),i} \right)^2}{2\sigma_i(\omega-1)} \tag{7.8}$$

and where ω is the size of a temporal window over which matching from t backwards to $(t-\omega-1)$ occurs, σ_i are estimates of the standard deviation for each of the trajectories i for the whole sequence. The term $\alpha m^{(\mu)}_{(\phi-\rho j),i}$ performs the dynamic time warping of the trajectory i from the training data. The model number is μ, the trajectory is shifted by ϕ, interpolated by ρ, and scaled by α. Hence Equation (7.8) represents the mean distance between the test gesture and a DTW'd model for ω-sized window of the trajectories.

Using S values of $p(\mathbf{z}_t \mid \mathbf{s}_t)$, it is possible to create a probability distribution of the whole search space at one time instant. Each conditional probability acts as a weighting for its corresponding state and with successive iterations, the distributions of the states in the search space cluster round areas which represent the more likely gestures. The weighting or normalised probabilities are calculated as follows:

$$\pi_t^{(n)} = \frac{p(\mathbf{z}_t \mid \mathbf{s}_t^{(n)})}{\sum_{i=1}^{S} p(\mathbf{z}_t \mid \mathbf{s}_t^{(i)})} \tag{7.9}$$

From these weights, it is possible to predict the probability distribution over the search space at the next time instant. Thus, more probable states are more likely to be propagated over the total time of the observed sequence. It is emphasised here that more than one probable state can be propagated at each time instant.

The sample set is first initialised by sampling uniformly for each parameter for every $\{\mathbf{s}_t^{(n)}, n = 1, ..., S\}$ samples,

$$\mu \in [0, \mu_{max}] \tag{7.10}$$

$$\phi = \frac{1-\sqrt{y}}{\sqrt{y}}, \tag{7.11}$$

$$\alpha \in [\alpha_{min}, \alpha_{max}] \tag{7.12}$$

$$\rho \in [\rho_{min}, \rho_{max}] \tag{7.13}$$

where $y \in [0, 1]$ and the initial value of ϕ has been defined such that the phase shift is minimised for bootstrapping purposes. The maximima and minima for each of the parameters was defined similarly to Black and Jepson's paper such that $/mu_{max} = 6$, $\alpha_{max} = 1.3$,$\alpha_{min} = 0.7$, $\rho_{max} = 1.3$, and$\rho_{min} = 0.7$.

S samples are initialised, where S is chosen to be 1000. Once the samples have been initialised, the states to be propagated must be chosen. This is done by constructing a cumulative probability distribution using the weights $\pi_{t-1}^{(n)}$, as shown in Figure 7.3. A value r is chosen uniformly and then the smallest value of the cumulative weight, $c_{t-1}^{(n)}$ is chosen such that $c_{t-1}^{(n)} > r$, where $(t-1)$ represents the current time frame and t indexes the samples and weights of the next time frame that is being predicted. The corresponding state $\mathbf{s}_t$ is then selected for propagation. With this method of selection, larger weights are more likely to be chosen. The ordering of the cumulative weight distribution is therefore irrelevant. To avoid getting trapped in local minima or maxima, 5% to 10% of the sample set are randomly chosen and initialised, as described above.

After states have been selected for propagation, the parameters for that state, at the next time step are predicted using the following equations:

$$\mu_t = \mu_{t-1} \tag{7.14}$$

$$\phi_t = \phi_{t-1} + \rho + \mathcal{N}(\sigma_\phi) \tag{7.15}$$

$$\alpha_t = \alpha_{t-1} + \mathcal{N}(\sigma_\alpha) \tag{7.16}$$

$$\rho_t = \rho_{t-1} + \mathcal{N}(\sigma_\rho) \tag{7.17}$$

where $\mathcal{N}()$ is the normal distribution and σ_* is the standard deviation or uncertainty in the prediction. The values of σ_* was chosen to be 0.05 for ϕ, and 0.1 for α, and ρ. μ is not changed during this stage but can be altered if the state proves not to provide a likely match with the observation trajectories.

After this stage, the new state is evaluated using the probability $p(\mathbf{z}_t \mid \mathbf{s}_t^{(n)})$. If the conditional probability is zero then the state is predicted again using the above equations

and $p(\mathbf{z}_t \mid \mathbf{s}_t^{(n)})$ is recalculated. If this process needs to be repeated more than a predetermined number of times, then, the state is deemed unlikely and it is reinitialised using the random initialisation described previously. The number of 'tries' is a pre-determined amount and this was chosen to be 100. Once all S new states $\{\mathbf{s}_t^{(n)}, n = 1, ...S\}$ have been generated, the normalised weights, $\pi_t^{(n)}$ are recalculated for state selection and propagation at the next time instant. The process of selection, prediction and update is repeated until the end of the observed sequence or the gesture is considered recognised is reached. The recognition criteria are described in the following section.

7.1.3 Experiment

The Condensation algorithm was implemented using MATLAB to recognise six possible gestures, as shown in Figure 7.1. Video images of these six gestures were taken from 16 different subjects, four times. The gestures were ; 'come', 'go', a high wave, a low wave, point left, and point right. All subjects were asked to perform the 'come' and 'go' gestures with both hands and the waving pointing gestures with their right hand. nine subjects were chosen for input observation sequences and six of these nine were used for model training, while all nine were used for testing.

The model was trained using 6 subjects, chosen at random, to represent the model data or training set for the algorithm. The model trajectories were created by interpolating each example of a particular gesture, to the mean length of the trajectory for that gesture. Then, the mean value at each time step, for each of the four trajectories, was calculated. Hence, each gesture was represented by four model trajectories corresponding to motion area, x and y displacement and elongation. These were then used to match with the observed gestures.

The video sequences that were used were manually trimmed, so that each sequence contained one gesture, performed once. However, despite this, there were still some variations in the onset of motion. Therefore, the likelihood of a test gesture matching a particular class was determined by using a thresholded ratio of the first and second highest likelihoods from all the test classes. If the ratio was high enough for a sustained period, this meant that the detected class was suitably discriminative from all others. However, if the ratio fells below the

Gesture	Chris	Cth	Dennisp	Djones	Kate	Paulv	Simon	Sylvia	Tom	TPR
Come	0	1	0.25	0	0.5	0	0.75	0.25	0.25	0.333
Go	0	0	0.75	1	0.75	0.25	0.5	0	0	0.361
Wavehi	0.25	0	0	0	0	0	0	0	0	0.0278
Wavelo	0.75	1	0.5	0	1	1	1	1	1	0.806
Pntrt	0	0	0	0.75	0.75	0	0	0.25	0	0.194
Pntlft	0.25	0	0	0.25	0.75	0	0	0.75	0	0.222

Table 7.1: Summary of the TPR of each gesture for each test subject. The last column shows the mean TPR
[htb]

threshold, for a certain number of consecutive frames, then the states would be reset and the gesture, considered recognised. As well as this, a maximum value for the phase parameter ϕ was used so that when ϕ was greater than ϕ_{max}, the gesture would also be considered recognised. This was used to emulate a situation where the end of a gesture might be unknown. The algorithm stops when either one of these 2 criteria are satisfied.

The overall performance of the algorithm was not good. Most subjects only had 2 or 3 gestures that were recognised at all. Only one subject had 5 out of 6 gestures recognised once or more. The true positive rates (TPR), false positive rates (FPR) and accuracy for each gesture for every subject is shown in Tables 7.1, 7.2, and 7.3.

Gesture	Chris	Cth	Dennisp	Djones	Kate	Paulv	Simon	Sylvia	Tom	FPR
Come	0	0.5	0.375	0	0	0	0	0.2	0.429	0.171
Go	0.167	0	0.5	0.67	0.333	0.2	0.125	0	0	0.260
Wavehi	0.429	0.111	0.143	0	0	0.167	0.181	0	0.167	0.115
Wavelo	0.867	0.714	0.636	0.273	0.083	0.923	0.706	0.5	0.917	0.643
Pntrt	0.286	0.111	0.4	0.444	0.143	0.5	0	0	0.444	0.259
Pntlft	0	0	0	0.125	0	0	0	0.571	0	0.127

Table 7.2: Summary of the FPR of each gesture for each test subject. The last column shows the mean FPR

'Kate"s gestures were the easiest to identify out of all the subjects, with the highest number of correct predictions, and relatively low false positive rates for her gestures. This was because her gestures tended to be most similar to those of the subjects in the training set.

The gestures of 'Chris' ,'Paulv' and 'Tom' performed the worst out of all the subjects. On inspection of the actual gestures of these subjects, it was clear that their gestures tended to be

Gesture	Chris	Cth	Dennisp	Djones	Kate	Paulv	Simon	Sylvia	Tom	Mean
Come	0.556	0.667	0.5	0.667	0.882	0.556	0.9	0.643	0.455	0.660
Go	0.5	0.667	0.6	0.5	0.681	0.5556	0.75	0.692	0.556	0.619
Wavehi	0.455	0.618	0.545	.667	0.789	0.5	0.6	0.692	0.5	0.614
Wavelo	0.263	0.444	0.4	0.533	0.938	0.294	0.429	0.643	0.313	0.464
Pntrt	0.455	0.615	0.429	0.615	0.833	0.357	0.692	0.75	0.385	0.579
Pntlft	0.625	0.667	0.6	0.667	0.938	0.556	0.692	0.5	0.556	0.654

Table 7.3: Summary of the accuracy of each gesture for each test subject.

less conventional than those that were used in the training set. For example, 'Chris" 'come' gesture involved just moving the 2 index fingers in cyclic motion (see Figure 7.4), whereas, on the whole, the subjects from the training data used the whole hand and sometimes the arms as well. Hence, the centroid of motion would have described much smaller arcs for 'Chris" gesture compared to those in the training set.

The results from studying the performance of individual subjects proves that the algorithm is good at identifying gestures which are similar to those in the training data. However, this does imply that recognition is very dependent on choosing the right set of training examples for the algorithm to match input trajectories. It also indicates that the size of window and the warp ranges for the state parameters could have a significant effect on the results.

Overall, the algorithm found it easiest to recognise the low waving gesture. However, although it had by far the highest TPR, it also had the highest FPR. Another interesting observation was that the high wave had the lowest TPR and poor FPR. The confusion matrix shows that the high wave was almost always mistaken for a low wave. The features were extracted based on the displacement of the centroid of motion. These 2 gestures were often mistaken for the other since the upward motion of the hand tended to be very fast and hence was captured by very few frames. This meant that even small variations in the state ϕ could lead to mis-classification. Often, the gesture was identified before the subject lowered his or her hand back to the 'rest' position and hence a high and low wave would be almost indistinguishable. Looking at the actual trajectories in Figure 7.5 where part (a) shows the feature trajectories for the high wave and (b) shows those for the low wave, the 2 gestures are very similar except for the much higher motion areas at the start and end of the high

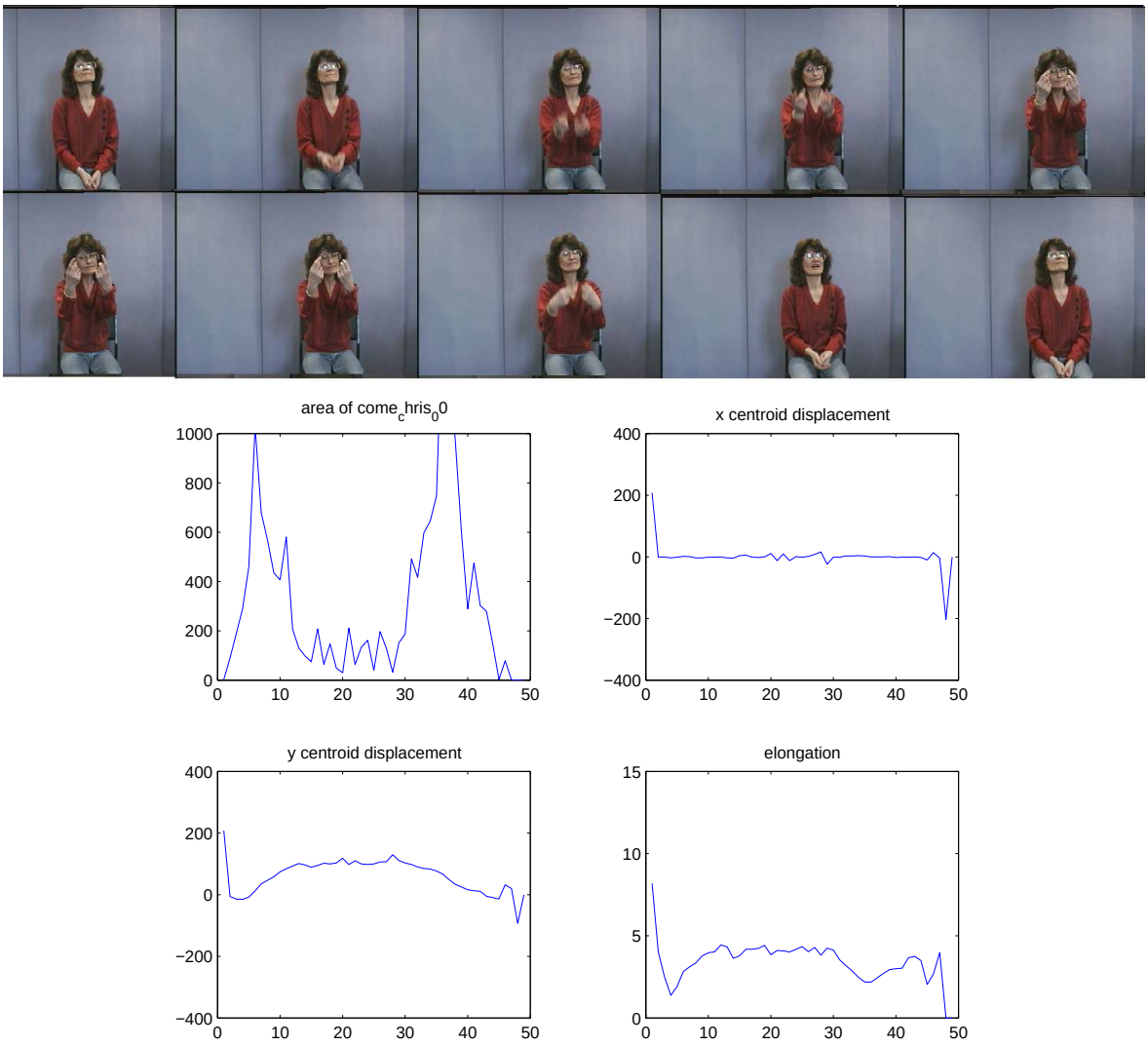


Figure 7.4: Feature vectors of the ‘come’ gesture from the subject, ‘Chris’.

waving gesture. Given enough warping, it would probably be possible to mistake the high wave for the low wave. This brings into question the validity and sufficiency of the feature set that was used.

The results show that the actual feature extraction and representation technique had a great effect on the performance of the algorithm. In fact, the algorithm actually performed well, given the simplicity of the trajectories that were used to represent the gestures. Perhaps if more features were included in the feature set, such as the actual centroid of spatially homogeneous areas of motion, it might have increased the discriminance between the gestures. However, using just 2-frame temporal differencing identifies the pixels that exhibit high intensity changes but this does not necessarily identify spatially meaningful homogeneous regions of change that can be equated to some higher level semantics. More sophisticated forms of temporal frame differencing will be used later in Chapter 5 where experiments show that even applying an adaptive background model doesn't always lead to an accurate representation of the background and foreground.

As described at the beginning of this report, one of the difficulties of gesture recognition is the variable nature of the same gestures from person to person, and also from the same person. To a certain extent, the algorithm was able to deal with identifying the same gesture from different people. However, it was very rare that many examples of the same gesture from one subject were always recognised correctly. Also, the more unique gestures of some subjects were not recognised at all if they were slightly different to the general method used in the training data. An example of this would be 'Chris"s 'come' gesture, which involved using just the index fingers rather than the whole hand for the cyclic motion in the middle of the gesture (see Figure 7.6). This implies that if more training data was used, it is likely that we could have captured more examples of different nuances of the same gesture.

Discussions and Conclusions

We can see from this experiment that even with just 6 gestures to classify, the results were poor. However, it was not the Condenation algorithm that was at fault. We have seen that when gestures were not recognised, it was because the feature set did not represent the

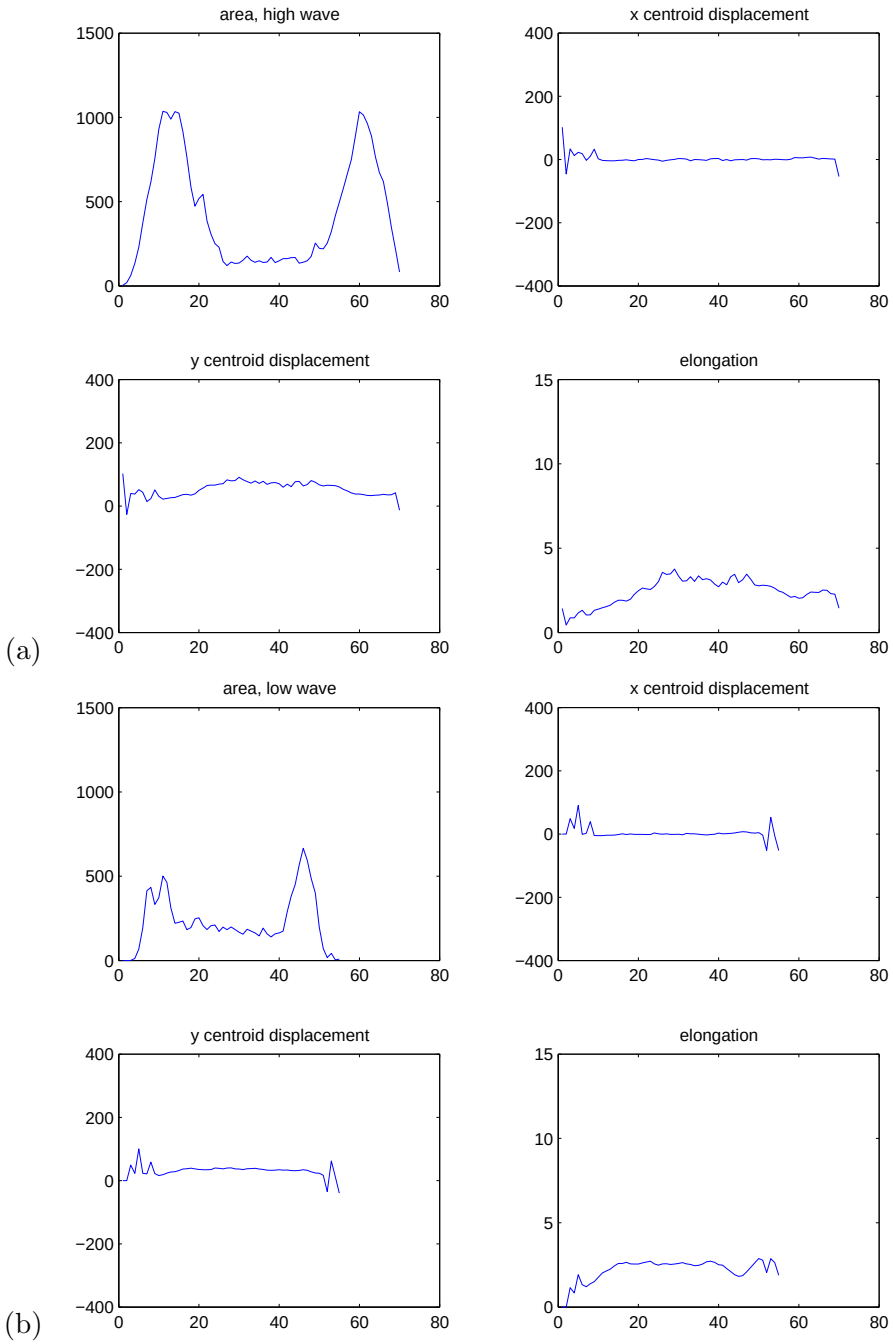


Figure 7.5: Feature vectors of the high waving gesture (a), and low waving gesture (b) from the training set.

(a)

(b)

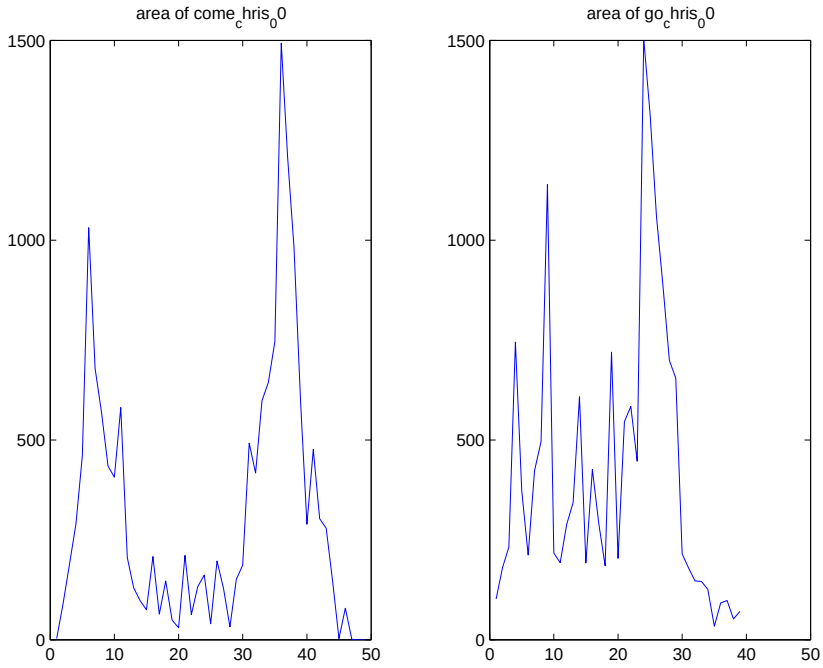


Figure 7.6: Comparison of the motion area trajectories of the (a) 'come' and (b)'go' gesture for subject, 'Chris'.

gesture in a suitably discriminative manner. For example, the method relied on all motion to originate from a single object. So this did not cater for the possibility that gestures could be performed by 2 arms or that different parts of the body had to be articulated in order to perform the gesture.

In terms of the classification process, a representation of all gestures in the set was formed from averaging across each feature trajectory for each subject. This is clearly impractical because even if one person performs a gesture in a slightly different manner to the majority of people this does not necessarily invalidate the gesture. So even the generalisation technique of local warping was insufficient. A more sensible classification technique might involve grouping gestures into meaningful clusters and subsets in order accurately represent their variable nature. It was also apparent that treating all the object motion as one entity was unrealistic. Therefore breaking up the motion into smaller components seems sensible since the human form is naturally deformable.

7.1.4 Searching for a Better Representation

The previous section identifies a clear need to find a more effective method of representing the features in the video sequences. It was apparent from the gesture set that even the orientation of the motion area as well as the elongation may have provided a more discriminative representation. For example, the low wave and pointing gesture may have not been as easily misclassified since the orientation of the motion areas of both are clearly very different. It would also be useful if the motion could be divided into subparts to preserve spatial ordering.

Multidimensional receptive field histograms (MDRFH) were first used for object recognition by Schiele and Crowley (1996). This technique created a set of data to represent an object at a number of positions and image-plane orientations. Hence the method was invariant to these parameters. This idea was carried further by Chomat et al. (2000) to recognise activities from human motion. There are a few advantages of using receptive fields since it must inherently reduce an image into smaller 'fields' which satisfies our need for spatially separated motion descriptors. The idea of using orientation filters to perceive motion was presented by Adelson and Bergen (1985). They proposed perception of motion through applying orientation filters

to spatio-temporal slices. Using orientation filters to extract information at each frame of an image sequence is close to the biological method with which the eye extracts visual information (Doll and Home, 2001). For this reason, it would be interesting to see the effect of using multi-dimensional receptive field histograms on a set of raw data. The following section provides a description of this method of feature extraction.

The idea behind MDRFH is to provide a rich representation of movement through accumulation of spatial orientation information over time. This is achieved through the use of multi-orientation filters to represent each frame of an image sequence. Statistics about a particular intensity value of each pixel are accumulated to provide a coarse probability density function or histogram. Steerable filters are used to compute orientations of the image sequence at each frame. Designing steerable filters is a very complex process and full details are provided in Freeman and Adelson's paper on the design and use of steerable filters (Freeman and Adelson, 1991). A brief description of first order steerable filters is provided in the Appendices.

In order to create multi-dimensional histograms of a sequence, many instances of the same sequence must be processed, to create statistics at each pixel of the image frame. These statistics take the form of probability density functions for each pixel of the image. Since it is impossible to generate a perfect probability density function, usually the function is approximated by a histogram, of an arbitrary number of bins.

The temporal aspect of the image sequence can be represented by accumulating the histogram over a window of time, or over a whole sequence. The process of creating a MDRFH at one scale from a sequence of images is illustrated in Figure 7.7.

Study of the MDRFH of gesture sequences

In order to understand a more descriptive method of feature representation, MDRFH was applied to the gesture data described in the previous section. Some adjustments were made to Chomat and Crowley's initial idea to reduce computational complexity. Firstly, only one scale of filter, with 4 orientations, was applied to each frame of a sequence. Secondly, it was decided that the temporal ordering of the sequence would be ignored in an attempt to

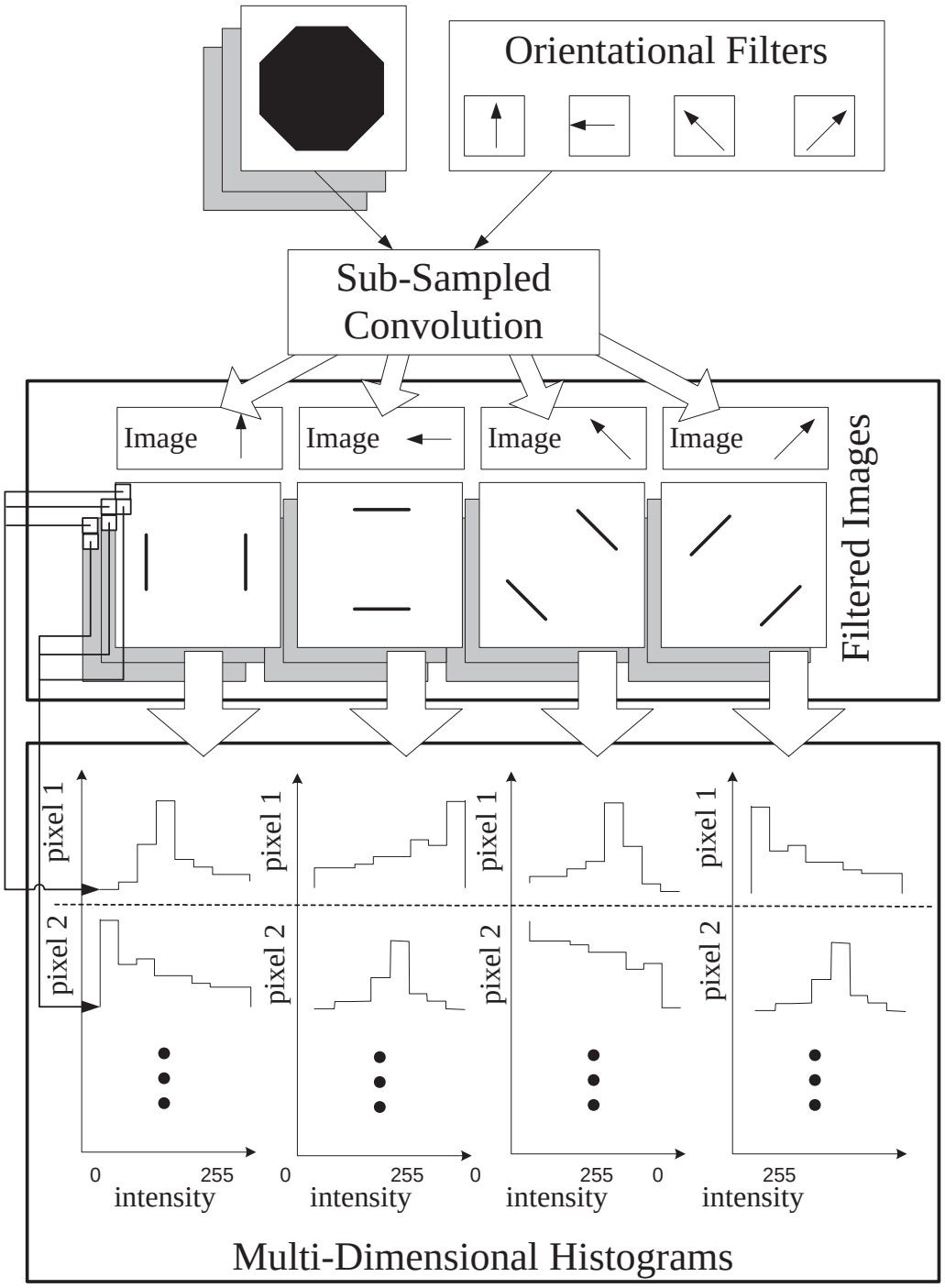


Figure 7.7: Flow diagram of applying the MDRFH on a sequence of images.

discover patterns in the multidimensional histograms from accumulating information about each pixel over the whole sequence.

Since each frame of the gesture sequence was very large (240×320), subsampled convolution was performed so that the resultant filtered images were only 11 × 15. This still required computation over 165 pixels but this was a suitable reduction in size, where approximately one pixel was the size of a hand. The 4 orientations for some frames of a high waving and pointing gesture are shown in Figure 7.8 and 7.9. These filtered and reduced versions of each frame provide a very course edge detection at different orientations. However, if the intensities at each of the cell locations are accumulated over, just 2 consecutive time frames, it is immediately obvious that some form of motion can be detected.

Experiment with PCA and MDRFH on gestures

A set of four subjects performing six gestures four times were chosen to populate the training data. For each grid cell of the resultant subsampled and filtered images, a histogram was created from the distribution of each of the corresponding filter responses across the each pixel over time. The possible intensity values were split into eight possible intervals and a histogram was accumulated for these intensity ranges over the whole gesture sequence (which had been manually segmented); there appear to be more similarities in the histograms, than differences. Hence, 165 (11 × 15) histograms were created for a particular class.

In order to represent multidimensional histograms in a meaningful way, Chomat and Crowley Chomat and Crowley (1998) used Principal Component Analysis (PCA). This meant that it was possible to map high dimensional data to a linear set of data. In other words, using PCA, it was possible to reduce the feature vectors to their statistically, most significant parts. For the gesture set that we have been using, this should mean that the most significant parts of each set of training data for each class should have a distribution which is centred on very different areas in this high dimensional space. A description of PCA can be found in Appendix 7.2.

It was found that the 6^{th}, 7^{th} and 8^{th} principal components appeared to provide the most discrimination between gestures. From Figure 7.10, the distributions of the gestures do

Figure 7.8: Frames of a high waving sequence, filtered at 4 different orientations. Frames are ordered row wise.

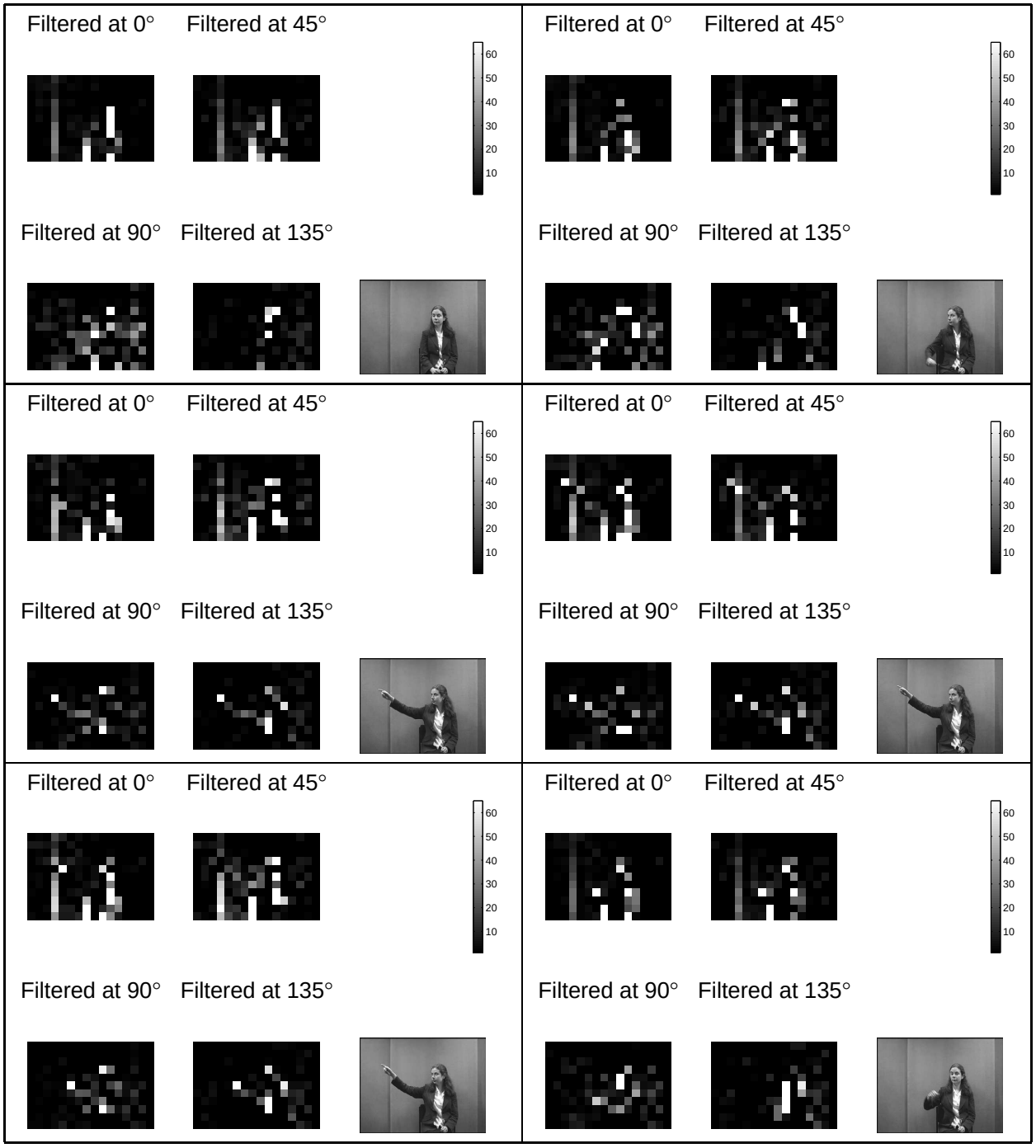


Figure 7.9: Frames of a pointing gesture, filtered at 4 different orientations. Frames are ordered row wise.

occupy different areas in high dimensional space. Furthermore, it appears that the problems with recognising high and low waves, encountered during implementation of the Condensation algorithm will provide no such hindrance in this case.

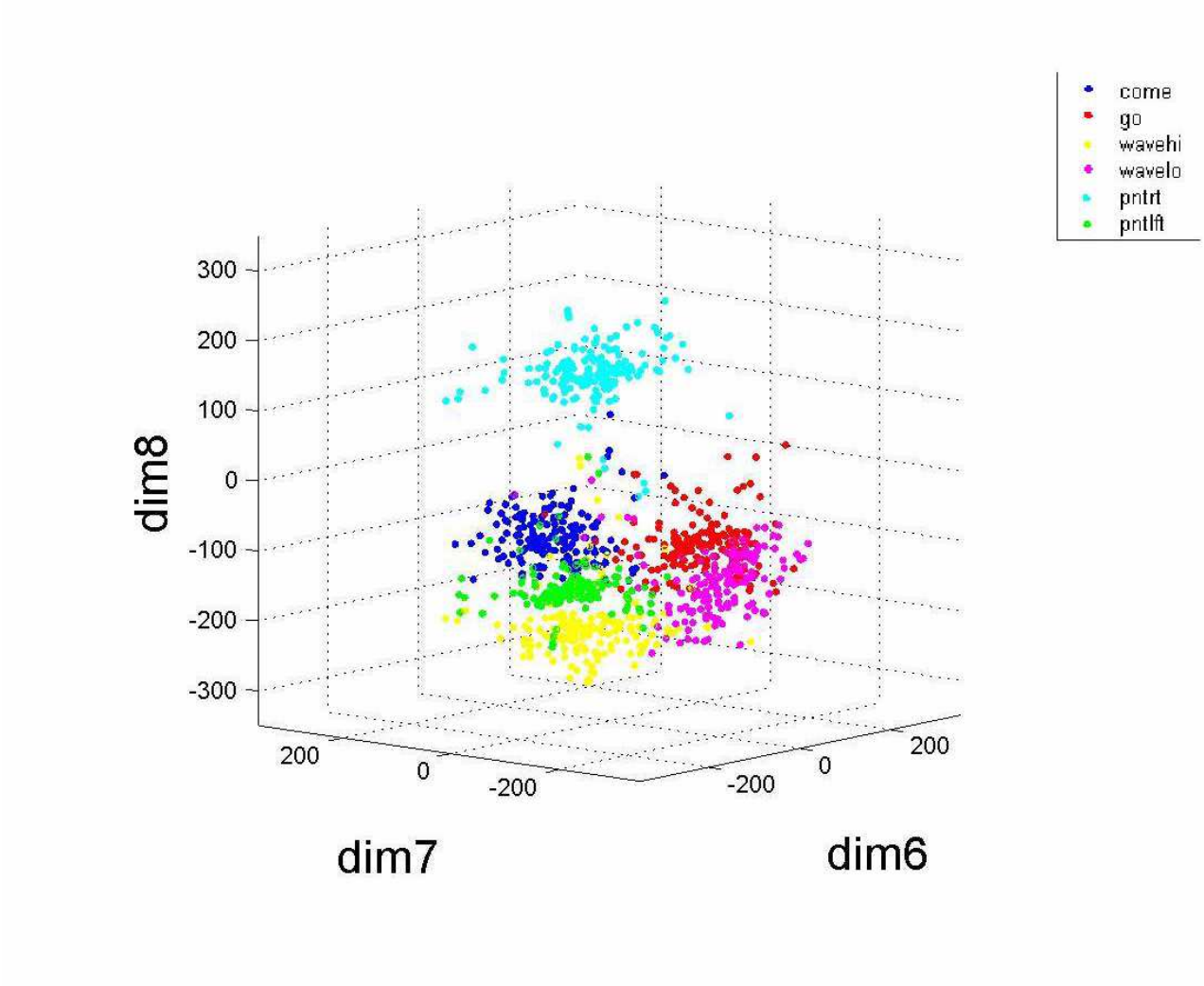


Figure 7.10: Graphs showing the 6^{th}, 7^{th} and 8^{th} principal components of the MDRFH data for all 6 gestures one a single plot

Whilst the experiment shown in this chapter illustrated the use of MDRFH accumulated over a whole gesture sequence, there are questions to be raised about how the representation would be affected by accumulating over smaller time windows. Furthermore, when the frames were subsampled to reduce computation, it was artificially set so that one pixel could still approximate the size of a hand. In more natural scenes, the size of hands or any other objects

of interest will not be known beforehand. Using the data itself in order to select the spatial scale seems a more sensible approach. This would also allow different spatial scales of moving objects to be used.

7.1.5 Conclusions and Summary

The experiments in this chapter have highlighted some of the issues that will be covered later. Firstly, effective feature extraction and representation of the data is extremely important for the success of the system. In particular, the sequences used for these experiments avoided many of the challenges that would be encountered with more cluttered indoor or outdoor scenes. The first method using two-frame temporal frame differencing lacked any sense of spatial grouping of the objects. Any pixel that differed enough from the background model was considered to be part of a single foreground blob, which made the assumption that all motion must come from the same object. The second method using MDRFH addressed this problem to a certain extent, by subdividing the frame into an equally distributed grid of orientationl receptive fields. The frame was also subsampled so that each grid location could, to some extent, represent something more semantically meaningful than a single pixel could. However, the degree of subsampling was determined by the size of the hand. From these observations, we can conclude that finding the correct spatial scale of moving objects is important. Using MDRFH also relied on extracting all orientation information from the scene. Given that each of the gestures was performed on a plain background, extracting edge orientations to represent meaningful cues was sufficient. However, in more realistic outdoor scenarios, many limitations are met since background clutter will also be extracted from the scenes.

A second issues that was highlighted was the use of either frame-based or template-based matching. In the first case, using single frame matching can lead to overfitting of the data since the context is much more defined by the matching of the test and model trajectories. Although the mean of the model trajectory was taken, which could imply greater generalisation, we can see from the results that there were unusual examples of the same class that were not represented at all. Using the template-based approach, and applying eigen value

decomposition to the data meant that each example of the model could be represented in a better segregated high-dimensional space, leading to better generalisation. The problem with the second method is that relying on template matching means that all sequential ordering of the gestures was discarded. This worked well for our particular examples since the gesture sequences were performed in a suitably simple fashion to facilitate minimal ambiguity between classes. A compromise between these two approaches will be investigated later.

7.2 Principal Component Analysis (PCA)

The aim of PCA and any other dimensionality reduction method, is to preserve as much relevant information as possible within a set of data whilst reducing the amount of information required to store the data. PCA is a linear transformation which maps a set of vectors $\mathbf{x}^n$ in d dimensional space to vectors $\mathbf{z}^n$ in an M dimensional space, where $M < d$. To aid transformation of the data, the vector $\mathbf{x}$ can be represented as a linear combination of a set of d orthonormal vectors $\mathbf{u}_i$[1].

$$\mathbf{x} = \sum_{i=1}^{d} z_i \mathbf{u}_i \tag{7.18}$$

Hence $\mathbf{u}_i$ performs the mapping of $\mathbf{z}$ to $\mathbf{x}$ and satisfies the orthonormality relation:

$$\mathbf{u}_i^T \mathbf{u}_j = \delta_{ij} \tag{7.19}$$

where δ_{ij} is the Kronecka Delta function. The vector $\mathbf{z}$ can be found by using the inverse of $\mathbf{u}_i$:

$$z_i = \mathbf{u}_i^T \mathbf{x} \tag{7.20}$$

Due to the orthonormal properties of $\mathbf{u}_i$ as seen in Equation 7.19,

$$\mathbf{u}^{-1} = \mathbf{u}^T. \tag{7.21}$$

[1] All notation is taken from Bishop (1995)

$\mathbf{u}_i$ can be regarded as a transformation that rotates the coordinate system of x to that of z. This can be proven by considering the eigenvector equation for a symmetric matrix.

$$\mathbf{A}\mathbf{u}_k = \lambda_k \mathbf{u}_k \tag{7.22}$$

where $\mathbf{A}$ is a $W \times W$ matrix and $k = 1, ..., W$. These k eigenvector equations can be written in matrix notation as

$$(\mathbf{A} - \mathbf{D})\mathbf{U} = 0 \tag{7.23}$$

where $\mathbf{D}$ is a diagonal matrix, with its elements set to the eigenvalues λ_k of $\mathbf{A}$

$$\mathbf{D} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_W \end{pmatrix} \tag{7.24}$$

and $\mathbf{U}$ is a matrix consisting of the eigenvectors $\mathbf{u}_k$ placed column-wise

$$\mathbf{U} = \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_W \\ \vdots & & \vdots \end{pmatrix} \tag{7.25}$$

Following from this, the solution of Equation 7.23 is singular if and only if the following equation holds:

$$|\mathbf{A} - \mathbf{D}| = 0 \tag{7.26}$$

Since $\mathbf{U}$ is orthogonal, matrix $\mathbf{A}$ can be diagonalised to obtain matrix $\mathbf{D}$, using Equation 7.23

$$\mathbf{U}^T \mathbf{A} \mathbf{U} = \mathbf{D} \tag{7.27}$$

Due to this orthogonal property of $\mathbf{U}$, it is possible to preserve the length of a vector when it is transformed. This can be proven by declaring Equation 7.20 in terms of lengths:

$$\|\mathbf{z}\|^2 = \mathbf{x}^T \mathbf{U} \mathbf{U}^T \mathbf{x} = \|\mathbf{x}\|^2 \tag{7.28}$$

Using Equation 7.20 again, it is possible to prove that the angle between the 2 vectors are also preserved.

$$\mathbf{z}_1^T \mathbf{z}_2 = \mathbf{x}_1^T \mathbf{U}\mathbf{U}^T \mathbf{x}_2 = \mathbf{x}_1^T \mathbf{x}_2 \tag{7.29}$$

Therefore, the matrix $\mathbf{U}^T$ maps one set of vectors to another by causing a rigid rotation of the coordinate system.

Once the matrix $\mathbf{U}$ has transformed the vector $\mathbf{x}$ to $\mathbf{z}$, it is possible to retain only a subset, M of the d basis vectors, $\mathbf{u}_i$. Hence, only M coefficients z_i are used, with the rest being replaced by constants b_i. The new approximation of $\mathbf{x}$ will be named $\tilde{\mathbf{x}}$.

$$\tilde{\mathbf{x}} = \sum_{i=1}^{M} z_i \mathbf{u}_i + \sum_{i=M+1}^{d} b_i \mathbf{u}_i. \tag{7.30}$$

The above equation can be used to approximate original vector $\mathbf{x}$ using the transformed vector z if M out of a possible N basis vectors $\mathbf{u}_i$ are chosen such that the error caused by the dimensionality reduction is minimised.

$$\mathbf{x}^n - \tilde{\mathbf{x}}^n = \sum_{i=M+1}^{d} (z_i^n - b_i)\mathbf{u}_i \tag{7.31}$$

Hence the equation to minimise for the sum of square errors for a whole data set of N samples is

$$E_M = \frac{1}{2} \sum_{n=1}^{N} \|\mathbf{x}^n - \tilde{\mathbf{x}}^n\|^2 = \frac{1}{2} \sum_{n=1}^{N} \sum_{i=M+1}^{d} (z_i^n - b_i)^2. \tag{7.32}$$

where due to the orthonormality relation of Equation 7.19, the the basis vectors do not contribute to the sum of square error. To find the minimum, we can differentiate E_M with respect to b_i and set the result to zero. Hence b_i can be found by using

$$b_i = \frac{1}{N} \sum_{n=1}^{N} z_i^n = \mathbf{u}_i^T \overline{\mathbf{x}}, \tag{7.33}$$

where $\overline{\mathbf{x}}$ is a mean vector and is defined as

$$\overline{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}^n. \tag{7.34}$$

Using Equation 7.20 and 7.33, it is possible to rewrite the sum of squares error in terms of the original set of vectors $\mathbf{x}^n$ as

$$E_M = \frac{1}{2}\sum_{i=M+1}^{d}\sum_{n=1}^{N}\mathbf{u}_i^T(\mathbf{x}^n - \overline{\mathbf{x}})^2 \tag{7.35}$$

$$= \frac{1}{2}\sum_{i=M+1}^{d}\mathbf{u}_i^T\mathbf{\Sigma}\mathbf{u}_i \tag{7.36}$$

where $\mathbf{\Sigma}$ describes the correlation between each $\mathbf{x}_i^n$ with every other $\mathbf{x}_j^n$. In other words, $\mathbf{\Sigma}$ is the covariance matrix of the set of vectors $\{\mathbf{x}^n\}$ and is defined as

$$\mathbf{\Sigma} = \sum_n(\mathbf{x}^n - \overline{\mathbf{x}})(\mathbf{x}^n - \overline{\mathbf{x}})^T. \tag{7.37}$$

E_M is minimised with respect to the choice of basis vectors $\mathbf{u}_i$ when

$$\mathbf{\Sigma}\mathbf{u}_i = \lambda_i\mathbf{u}_i \tag{7.38}$$

is satisfied. Hence, $\mathbf{u}_i$ are the orthonormal eigenvectors of the covariance matrix and λ_i are the eigenvalues. By substituting Equation 7.38 into Equation 7.2, we can define E_M in terms of the eigenvalues λ_i.

$$E_M = \frac{1}{2}\sum_{i=M+1}^{d}\lambda_i \tag{7.39}$$

Thus, minimising the E_M is simplified to discarding the $d-M$ eigenvalues of the covariance matrix $\mathbf{\Sigma}$ and hence their corresponding basis vectors. The remaining basis vectors $\mathbf{u}_i$ are called the principal components. That is to say, they describe the principal modes of variation of the d dimensional set of vectors $\mathbf{x}^n$, whilst the corresponding eigenvalues λ_i describe the magnitude of variation. Hence the transformed vectors $\mathbf{z}^n$ form a diagonal matrix, with its M elements set to the eigenvalues.

It is clear that using PCA, it would be possible to treat all the MDRFH data from the previous chapter within much smaller dimensions. Whilst it is likely that the data is non linear, PCA is simple to implement and will provide more insight into the general problems with feature representation. The next section will perform PCA on the MDRFH data generated in the previous chapter to see whether the feature extraction technique is suitable for discriminating between gestures.

7.3 First order steerable filters

A 2-dimensional circularly symmetric Gaussian function G, written in Cartesian coordinates x and y is expressed :

$$G(x, y) = e^{-(x^2+y^2)} \tag{7.40}$$

The first derivative of the Gaussian function, with 0° orientation, is written as:

$$G_1^{0^\circ} = \frac{\partial}{\partial x} e^{-(x^2+y^2)} = -2xe^{-(x^2+y^2)}. \tag{7.41}$$

If this same function was rotated by 90°, the Gaussian function would be expressed in terms of y instead of x:

$$G_1^{90^\circ} = \frac{\partial}{\partial y} e^{-(x^2+y^2)} = -2ye^{-(x^2+y^2)}. \tag{7.42}$$

These first order rotated versions of the Gaussian function are useful since they can be applied to images to detect edges at different orientations. They are the basis filters for G_1^θ since it is possible to create first order orientation filters of any orientation from these filters , using the $sin(\theta)$ and $cos(\theta)$ terms to interpolate between them.

$$G_1^\theta = cos(\theta)G_1^{0^\circ} + sin(\theta)G_1^{90^\circ} \tag{7.43}$$

Using orientation filters, it is possible to extract edges of a particular orientation of an image, as seen in Figure 7.11. In this example, 4 different orientations have been used to describe the image of an octagon. The 4 orientations are 0°, 45°, 90°, 135°. Hence, it is

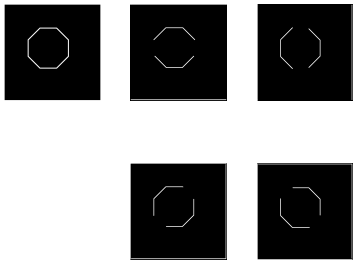


Figure 7.11: Different orientation filters applied to an image of an octagon. From top left clockwise, The image of the octagon, Results of the image filtered by first order Gaussian function with 0°, 90°,135°, and 45° orientations respectively.

possible to pick out 4 different edge orientations of the octagon. Since convolution requires heavy computation, it is usual to perform sub-sampled convolution on the image.