# Spotting the bad guys - A journey from surveillance video to automated crime fighting

Hayley Hung

# 1 Find a needle in a haystack

Following the events of 7 July 2005, it came to light that some of the suicide bombers were already known to the UK intelligence agents. What were the chances that they would all decide to meet on two different occasions within the space of a few months, at the same train station in London? Would it not have been useful to have some way of knowing when suspicious people decide to meet, just in case we could save many lives? The usual solution to this problem is to employ more people to monitor behaviour in public spaces. However, this is expensive and may not provide us with the best answer. If we had known where to look from the start, then many more crimes could have been prevented. Without this knowledge, finding surveillance footage of how the bombers planned their attack is like finding a needle in a haystack.

In the last decade, security cameras have become a common sight in the urban landscape. Being able to monitor many different places from one location has aided crime prevention greatly. In fact, just placing cameras in public spaces has become quite a deterrent for criminals. However, the numbers of security cameras that are being installed in everything from public streets to train carriages has surprisingly spawned a problem. On the London underground alone, there are at least 9000 security cameras. Security staff can have as many as 60 cameras to watch at any one time. Monitoring that many cameras is clearly an extremely difficult task, requiring large amounts of concentration for long periods of time. 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 by all these CCTV cameras. Automation is clearly the next step.

Simply trying to extract useful information from a scene is a difficult problem. If we think of a video as a sequence of many images, where each image is made up of a grid of pixels, we can begin to understand the problem of how to reduce this to something meaningful. How do we pick out the meaningful pixels? Then how do we group these pixels into meaningful regions? Once we've done that, how can we make sense of all what we have extracted?

In this article, we will explore the problems that must be faced when trying design automated surveillance systems and what state-of-the-art solutions are available.

# 2 The human visual system

The computer vision research community has drawn much inspiration from the way the human eye works. At every moment, our eyes are bombarded with so much visual information that we have evolved to make sense of it by simplifying things. This is formalised in Gestalt Psychology as the Gestalt Law of Minimum Principle [1]. We can illustrate this phenomenon in Figure 1. Here we see that although the picture contains separated shapes, we try to make sense of their formation so we see a white sphere. The problem of automatically grouping together or making

sense of many disparate parts is also referred to as the *binding problem*. It is something that occurs in the human visual system very easily but is clearly not easy to automate.

Figure 1: Images demonstrating how the human visual system groups parts of an image together. 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`.

So far, we have only talked about images but clearly the world we live in is made up of moving images. Furthermore, adding motion to the images we see allows us to understand it better. Going back to Gestalt theory, let's turn to some scientific results found by one of the early founders of this school of thought, Max Wertheimer. In 1912, Wertheimer discovered the Phi phenomenon [19]: humans can perceive motion, even if something is not physically moving. He found that when two lights, which were placed appart, were switched on alternately, instead of seeing two lights being turned on alternately, what was perceived was one light moving from side to side. This is also how an automated system would perceive motion. Whilst we see the world continuously, video cameras take snapshots of a scene in quick succession. The capture is fast enough that when we are shown these images, the pictures do seem to be moving. Automated surveillance systems rely on these image sequences to understand the scenes they are looking at (see Figure 2(D)).

Perception of motion is a really important part of how we understand the world. By 'joining the gaps' or *interpolating* between a series of static images, we are able to second guess where an object might move next. Second guessing, or *prediction*, is a really important part of how we understand the world since it allows us to separate out meaningful and less important information.

For example, if we are looking at a very busy scene of a market at 10am, many people are walking around. There is a lot of motion, but our eyes do not watch everything in the scene. We build a *model* of how we see the world at any one time. That means that we have prediction mechanisms that can determine where people are likely to walk or what sorts of clothes they are likely to wear. If suddenly someone runs through the market rather than walking, this really catches our attention. This is the idea behind research into automated abnormal behaviour

detection. We build a model of what we think the world is like. Then, if some behaviour does not fit this model, then it triggers alarm bells that something unusual has happened.

Making a model is not a trivial task. What we expect to see changes all the time so our model has to be different depending on circumstance. If we were to revisit the market at 1 am, 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. So *context* also plays a key role in how we build models of what we would consider to be normal. How we construct these models strongly affects the success rate of what we are trying to automatically detect.

So far we have seen how complex and sophisticated the human visual system is. Using this biological system to inspire an automated one seems promising..

# 3 From biology to autonomy

Figure 2 shows how an automated system can be inspired by human visual perception and also how they might differ. Let us investigate this by looking at how the steps involved in building an automated system might compare to the human visual system.

## 3.1 Feature extraction (Figure 2E):

This is the stage where regions of interest are extracted from the raw data or pixels of the image sequence. The only difference is that the human visual system extracts continuous features whereas the automated system extracts discrete features. These regions are grouped and then tracked. This is similar to the biological system, as shown in Figure 2A where region grouping is performed based on detecting features and finding similarities between them. Regions are then tracked in both cases.

## 3.2 Object recognition (Figure 2F):

This can be either recognition of object categories, scenes or people. The human visual system will also perform this function in order to simplify the surroundings, as shown in Figure 2B.

## 3.3 Behaviour analysis (Figure 2F):

Once the objects have been recognised, we can pick out which objects are of more interest to us, depending on the context. Behaviour analysis usually involves building models of how an object usually behaves so that we can for example, distinguish someone running from someone walking or standing. In the biological system, identifying behaviour is performed by using our experience, as shown in Figure 2B. Most automated systems rely on the models being trained

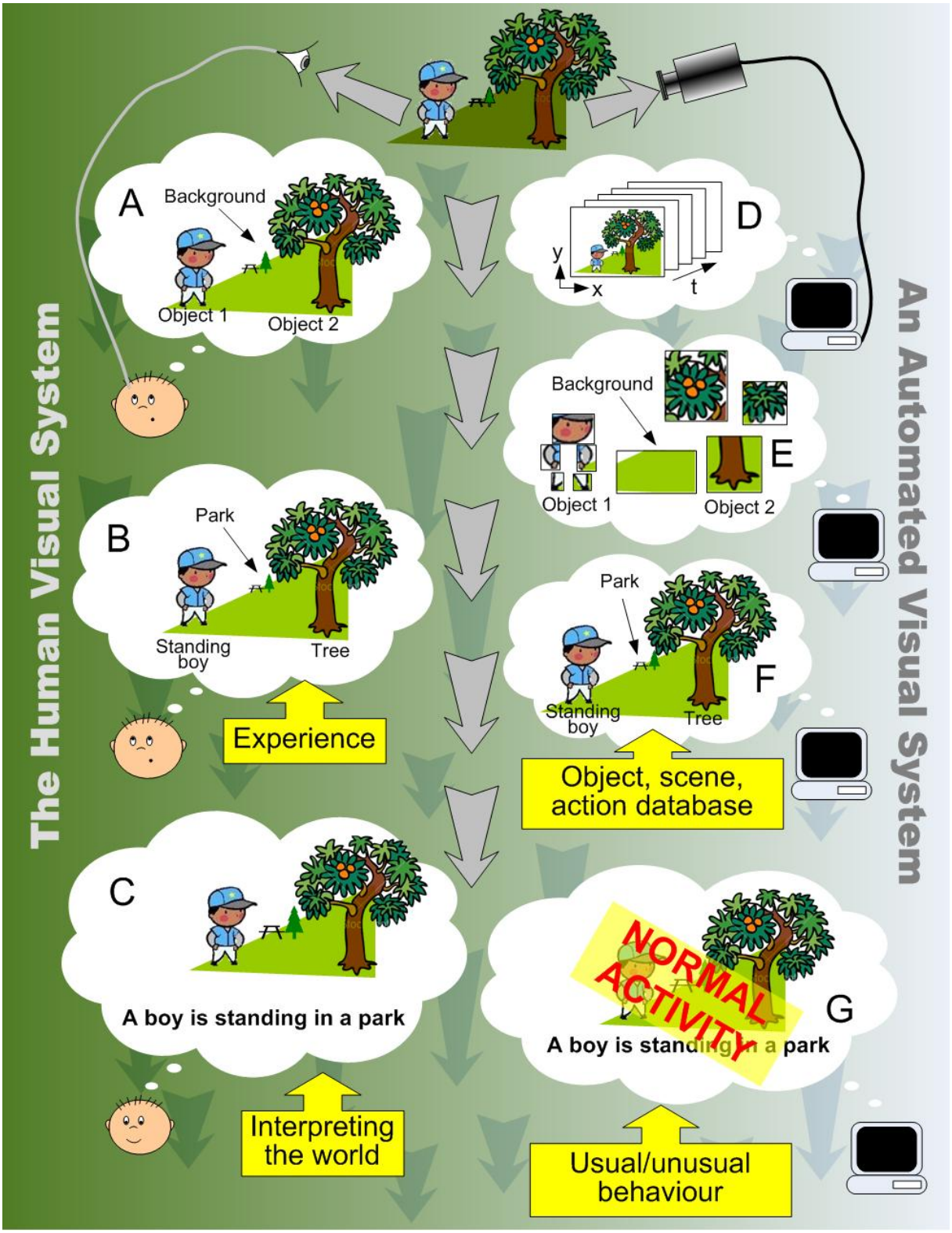


Figure 2: The similarities and differences between the human visual system and an automated system.

beforehand, though there are some models that are starting to become adaptive to learning models for different contexts.

For the human visual system it is essential to identify the objects before trying to analyse their behaviour. However, for automated systems, we can work straight from the extracted features. This will be discussed later.

### 3.4 Unusual behaviour detection ((Figure 2G):

Detecting unusual behaviour can be challenging because it does not happen often so modelling it is difficult. However, this part is crucial if we want to be able to detect suspicious behaviour or behaviour that does not fit with our model. This part is not essential to the human visual system but we still try and provide some way of simplifying the world around us by grouping the objects we see into one word or association. For example, if we see many trees and grass, we do not think of them as many separate objects. We automatically associate them with on expression, such as ”park”, as shown in Figure 2C.

## 4 Object recognition or identification

Detecting objects and being able to categorise them is useful for context modelling. If we can identify objects associated with different scenes, we can determine the location,the time of day or the type of action someone might be performing. The other more common form of object recognition is for face identification. If we can recognise a face, finding criminals who are still at large would be much easier.

Before we even start finding faces, we need to start looking for people first. Currently, the most popular object detector was devised by Viola and Jones [18]. Their method uses many features that may be considered to have a weak resemblance to the object we want to detect so that none on their own would be truly representative of it. However, when enough of these weak features are considered together, and we select those that work best in combination, they can have real discriminative power. The advantage of this particular detector is that although it can take months to train the system just to detect faces, when it comes to using it on real video sequences, the recognition times are extremely fast.

Once a face is detected, face identification is desirable, particularly for terrorist threats. Li et al. [13] devised a system that could recognise faces by creating surface identities or a fingerprint for the face. These were created by firstly taking pictures of the face at different angles and merging them into one image. This data from the images was then compressed into surface identities.

So it seems that the problem of face identification has been solved. However, this is not

(a) (b)

Figure 3: Results reproduced from [22] of the person detection system in a busy tube station. The images show the results without using the motion information to detect the people (a) and the system when just using person detector , using the motion information(b).

the case. Whilst small scale face recognition systems have shown a high accuracy, as yet, there are not successful implementations of commercial software for environments such as airports. Even if methods boast a 99% success rate, if the system is trying to detect one face in one thousand people, there is still a large number of people who might be falsly detected. Adding in other cues might help but with crowded public spaces, finding other ways of measuring identity is a challenging problem and remains to be solved.

Detecting people is a slightly more challenging problem than detecting faces since it is difficult to decide where different limbs of a person are going to be whereas the relative location of the eyes, nose and mouth are fairly consistent for faces. The problem with using weak features to detect people is that the system can sometimes mistakenly detect something to be a person which is not. This can be quite costly from a computational point of view since the system might want to track any objects that are detected as people. If too many incorrect results are found, it is easy to see that keeping track of so many objects becomes an impossible task. Recent research has explored using the motion information in the scene to determine whether it is worth trying to identify whether the moving area contains a person [22]. The results are shown in Figure 3 where (a) shows the results of a people detector without using motion information as a form of context, and (b) shows the results when motion information is also used where all the false detections have been removed.

From Figure 3, it is already apparent that performing face recognition on the detected people is quite difficult since the faces are so small. Low resolution issues are usually caused by the physical constraints of the video camera, which dictates that their position is fixed and the angle of the field of view is limited. Often, in order to maximise on the coverage area of a scene, the people tend to be very small and captured at low resolution. From these low resolution problems the research field of super resolution was born.

Given two neighbouring pixels in an image, what if we wanted to know the intensity value in between those two pixels? A simple solution would be to find the average intensity of the two neighbouring pixels. However, we are not gaining any new information from doing this. We are still reusing old information. The idea behind super resolution is that we can add our own expertise of the world to this guess [11]. Some of the newest results for face recognition and super resolution enhancement for face identification are shown in Figure 4. The enlarged faces are highly pixelated but with the super resolution technique, it is possible to add more definition to the contours of the faces and even add details around the eyes.

Figure 4: Super resolution and face detection. All the faces in the group photograph were detected. Two of them have been enlarged to show the original image and the enhancements using super resolution [11].

# 5 Classifying behaviour

Let's return back to our example of the market stall. What if someone walks into the market and starts punching someone else? This is clearly criminal behaviour but how can that be distinguished from someone waving to a friend or any other type of action? Being able to recognise different actions could help us to find suspicious behaviour or to predict when antisocial behaviour is about to occur.

Early methods of classification looked at motion patterns from blocks of pixels over time [2, 10, 20]. The disadvantage of these methods was that each pixel block was fixed so if the same action was performed in a different place, it would be impossible to identify the actions to be the same. Furthermore, these experiments took place in indoor scenes where each person was represented by quite a few pixels so there was much information about their motion. For surveillance data, however, it is likely that most people in the video will be far away from the camera and captured with relatively few pixels.

Recently, more sophisticated techniques have been designed so that actions can be recognised at low resolution, regardless of where a person is [14, 6, 5, 3]. Efros et al. [4] devised a technique for recognising the actions of footballers taken from TV coverage of a football match. The resolution of each footballer taken from a wide angle shot is very low. Each footballer is represented by about 20 pixels! They used some very detailed models to make the system work by detecting the motion of each part of the footballer's body and turning it into a 3D skeleton. Figure 5 shows some of the results. From the 3D skeleton that was generated from the low resolution image, a direct match for a person at a higher resolution could be found for each angle the person moved in. Whilst this technique showed great results, the inevitable disadvantage is the highly detailed models that are required to represent a person performing any action at many different view points. For a football match, there are at least a relatively limited number of actions that can be performed. However, in public spaces, many more different types of actions could be performed and even interactions between people can occur over much larger distances than what you would expect for a tackle in football.

Figure 5: Action recognition at a distance. The first row shows a zoomed in shot of the footballer, which is clearly represented at really low resolution. The second row shows the 2D skeleton of the footballer taken by taking motion from the images in the top row and learned models of where the joints in the person are likely to be. The third row shows the 3D skeletonisation of the footballer by taking motion information extracted from the first row and the 2D skeletons of the second row.

# 6 Unusual behaviour detection

Detecting unusual behaviour is a useful way of reducing large amounts of video to something manageable for security staff. There have been many lines of research into this area. Some have concentrated on recognising human actions or activities [9, 8, 23], others on the paths (or trajectories) that people or cars might take [12, 24, 15], or on more collective interactive behaviour [7, 5].

Once a moving region has been found, we want to start tracking it, to see if it takes a

predictable path or something more unpredictable. Tracking individual moving objects is not particularly difficult. However, when scenes start to get cluttered, then objects become hidden or occluded at different times and keeping track of where they are can be a problem. Fortunately, there are ways to solve this by combining knowledge of the direction an object is moving in, the colour information and configuration of the coloured regions. Figure 6 shows some tracking results from Zhou and Tao [24], showing two people walking past each other. Tracking them at first is easy since they are quite far appart from each other. However, as they get closer together, it becomes more difficult to decide which box corresponds to which person. It is even more difficult when one person hides the other.

Figure 6: Tracking two people who walk past each other. Results are taken from [24]

Returning to the idea of context, Vaswani et al. recently designed a system to detect abnormal behaviour in an airport where the context was determined by whether they were trying to track the motions of people or vehicles. From this, they were able to build models of normal people behaviour and vehicle behaviour [17]. Figure 7 shows some of the results of modelling people and vehicle motion from around an airplane. The clear circles show the normal paths taken by the people and the solid circles show that of the vehicles. Clearly, when a vehicle strays into the normal path of the people, then a trigger can be set off to alert security staff to be more vigilant of that particular event.

Perhaps the biggest question involving unusual behaviour detection is whether it is possible to predict antisocial behaviour before it happens. Pyschophysical studies [16] carried out at Sussex University showed that humans can do this. Subjects were shown 20 second video clips of surveillance data taken late at night. In each clip, those with true criminal or antisocial behaviour were stopped before the events had occurred. Subjects were asked to predict whether something bad would happen. In most cases, their predictions were correct.

So far, we have just made the assumption that making a model of normal behaviour was already assumed. However, just deciding on how to make the model has many pitfalls. For example, do we make a model and use it for all cases? This would not work for the market example where models of usual and unusual behaviour changes at different times of the day. Furthermore, how easy is it to train a system to learn a model? How long should we train a system to learn the patterns of activity from a busy airport? One day? One month? There are clearly many issues just to do with how this information should be learned. It seems logical

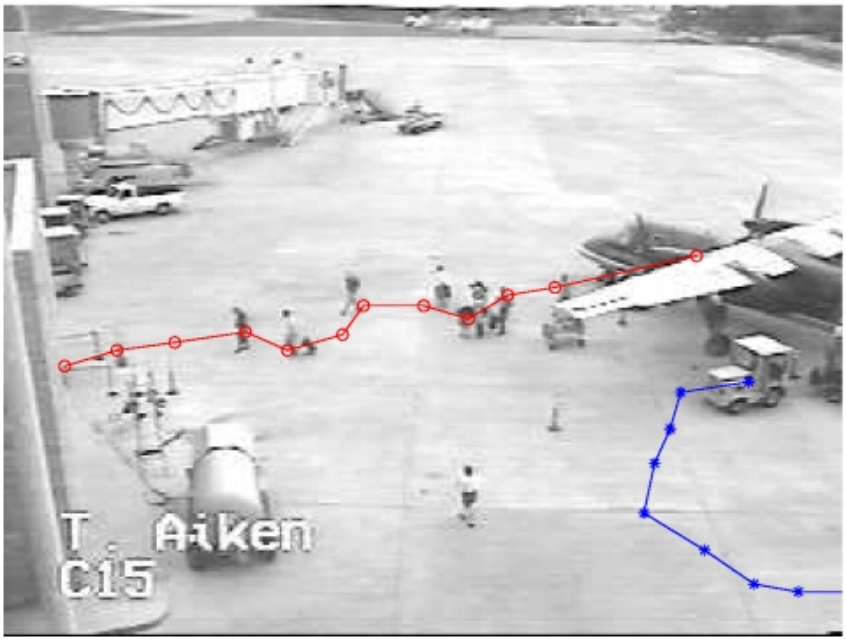


Figure 7: Models of normal behaviour are shown for people and vehicles with clear and solid circles respectively. Results are taken from [17]

therefore for a system to be adaptive. Even if the system has not seen a particular type of activity before, it can still build a model around it if it happens often enough over a period of time. Then, the learned model can become part of the normal behaviour. Xiang and Gong devised an algorithm to do just this, learning on-the-fly with an incremental model construction where the model constantly adapts depending on the context [21].

# 7 Ethical issues

The biggest moral objection to automated surveillance is the fear that personal privacy is forfeited. Companies selling automated surveillance software are aware of issues of privacy and have already taken steps to ensure that their systems are as unintrusive as possible. In fact, the automated system of SafeHouse Technologies is able to black out areas of a scene, which security staff are monitoring. Only areas of a scene that exhibit unusual behaviour are shown. This way, unless you are behaving suspiciously, no one need know anything about your private life. In fact, systems that only show areas of unusual behaviour are also likely to work faster since computing time isn't wasted in watching areas that never have anything unusual happening.

# 8 Challenges for the future

This area of research is still at its early stages, being less than 10 years old. Research councils and government funding agencies continue to show an interest so the future looks promising. However, commercial sponsorship is still yet to increase to a significant level due, to some extent, to the expectation that these systems should work perfectly or at least have a false alarm rate which is not impractical.

Evaluating solutions to this problem effectively remains to be a challenge since scientists must train and test these systems with real video scenarios. Clearly obtaining these videos for research purposes directly conflicts with personal privacy issues. In addition, even if these videos were made available, there is still the need to hand-label thousands of hours of video so that the systemc can trained and evaluated by human judgements. This is something which funding agencies are not able or willing to fund directly. Meanwhile, many researchers solve this by using their own 'home-made' videos which tend to be biased towards unrealistic scenarios, taken in 'easier' environments which are less problematic for computer vision systems. The extrapolation of these solutions to more challenging video data is high risk and is approached by few and generally require cooperation from companies or the government [21] .

Certainly the effect of world events has made finding concrete solutions to automated video surveillance more of a hot topic. However, it is down to society to decide whether they are willing to give up their personal privacy to allow machines to protect our streets in the future.

# References


[1] John Benjafield. *The developmental point of view. A history of psychology*. Simon and Schuster Company, 1996.

[2] O Chomat, J Martin, and J. Crowley. A probabilistic sensor for the perception and recognition of activities. In *ECCV (2)*, pages 487–503, 2000.

[3] James Davis and Hui Gao. Recognizing human action efforts: An adaptive three-mode pca framework. In *International Conference on Computer Vision*, pages 13–16, October 2003.

[4] A Efros, A Berg, G Mori, and J Malik. Recognizing action at a distance. In *ICCV*, 2003.

[5] A Galata, A Cohn, D Magee, and D Hogg. Modelling interaction using learnt qualitative spatio-temporal relations and variable length markov models. In *ECAI*, 2002.

[6] 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.

[7] H Hung and S Gong. Detecting and quantifying unusual interactions by correlating salient motion. In *AVSS*, 2005.

[8] H Hung and S Gong. A bottom-up approach to quantifying saliency in video. *To appear in EURASIP Special Issue on Tracking in Crowded scenes.*, 2006.

[9] Hayley Hung and Shaogang Gong. Quantifying temporal saliency. In *BMVC*, 2004.

[10] Y Ivanov and A Bobick. Recognition of visual activites and interactions by stochastic parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(22):952–872, 2000.

[11] K Jia and S Gong. Multi-modal tensor face for simultaneous super-resolution and recognition. 2005.

[12] N. Johnson and D. Hogg. Learning the distribution of object trajectories for event recognition, 1996.

[13] Y. Li, S. Gong, and H. Liddell. Video-based online face recognition using identity surfaces, 2001.

[14] Sangho 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.

[15] Chris Stauffer and Eric Grimson. Learning patterns of activity using real-time tracking. *PAMI*, 22(8), 2000.

[16] 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.

[17] N Vaswani, A Chowdhury, and R Chellappa. Shape activity : A continuous state hmm for moving/deforming shapes with application to abnormal activity detection. 2005.

[18] P Viola and M Jones. Robust real-time object detection. *International Journal of Computer Vision*, 2002.

[19] Max Wertheimer. *Classics in Psychology-Experimental Studies on the Seeing of Motion*. US/Mountain, 1961.

[20] A Wilson and A Bobick. Realtime online adaptive gesture recognition. *MIT Media Lab Perceptual Computing Section Technical Report*, 505, 1999.

[21] T Xiang and S Gong. Incremental visual behaviour modelling. In *IEE International Conference on Computer Vision*, 2006.

[22] Z Zhang and S Gong. Beyond static detectors: A bayesian approach to fusing long-term motion with appearance for robust people detection in highly cluttered scenes. 2006.

[23] Hua Zhong, Jianbo Shi, and Mirko Visontai. Detecting unusual activity in video. In *CVPR*, 2004.

[24] Y Zhou and H Tao. A background layer model for object tracking through occlusion. 2003.