Detecting and Quantifying Unusual Interactions by Correlating Salient Motion

H Hung and S Gong
Department of Computer Science
Queen Mary University
London, E1 4NS

Abstract

A significant problem in scene interpretation is efficient bottom-up extraction and representation of salient features. In this paper, we address the problem of correlating salient motion at a spatio-temporal level and also across spatially separated regions since it is in the interactions that more sophisticated scene interpretation can be found. We show that it is possible to spatio-temporally locate and detect salient motion events and interactions in two contrasting scenarios using the same hierarchical co-occurrence framework. Thus generating a concise description of a dynamic scene from the sequence data alone. Results show it is possible to reduce a highly populated multi-dimensional co-occurrence matrix representing correlations between salient motion regions, to a one dimensional vector with clearly separable unusual activity. The results also show that the method inherently provides a quantifiable measure of the saliency of an interaction through its frequency of occurrence.

1. Introduction

Pre-attentive feature extraction is an important stage of scene analysis. Without a suitably discriminative representation of unusual (salient) motion, clustering features extracted from the raw data is largely ineffective. Intuitively, to maximise the saliency of extracted features, top-down information must be injected into the process. However, a more favourable method would minimise reliance on scene-specific prior knowledge by using context drawn from the raw data alone. It is therefore both necessary and attractive to develop a bottom-up model that is capable of drawing correlation between spatially separated but temporally correlated as well as spatio-temporally correlated events. Thus higher level inference of more globally salient information is possible.

Zhong et al [10] used document clustering techniques for detecting unusual activity in video. However, their method did not address issues of temporally correlated but spatially separated behaviour. Their approach also relies on bipartite co-clustering to group together common prototypes and corresponding video segments. We prove 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 initially extracted, noisy and arbitrarily thresholded functional responses from the imagery data [10, 8].

Whilst popular techniques for extracting salient visual features for images or video use orientation filters [7, 1, 5], such techniques tend to favour features that produce higher magnitude responses from a rather arbitrarily chosen set of basis functions. The advantage of the Kadir and Brady's scale saliency algorithm [3] is that it is able to asses 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. This algorithm was extended [2] to variations in entropy over multiple temporal scales for extracting contextually salient features from video. In this paper, ambiguity in the spatio-temporal location of salient features due to a two-sided temporal sampling kernel was removed by using a one-sided version to express more precisely, the spatio-temporal salient activity within a scene.

To detect salient events at a higher level of inference, a popular approach is to cluster features extracted from the raw data and then apply co-occurrence techniques to them [8, 10]. Stauffer and Grimson [8] modelled the background through tracking mechanisms. However, in highly cluttered scenes, where partial or total occlusion of an object occurs, it is not always viable to perform multiple object tracking. Catering for robust tracking under partial or total occlusion conditions requires pre-determined contextual assumptions about what the scene may contain. We show 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.

We propose that it is possible to detect and quantify salient motion events from accumulating the co-occurrence of saliency values within a local spatio-temporal neighbourhood. This method accumulates co-occurrences of atomic salient motion descriptors based on spatio-temporally interacting neighbouring grid responses. This facilitates detection of unusual or salient motion caused by an individual event, or more complex multiple cause-effect phenomenon from spatially separated but temporally correlated scene locations. Specifically, salient events caused by motion that

would not be considered particularly meaningful individually, may define a more sophisticated level of understanding when addressed together. We propose that this information is inherently measureable from the raw sequence data and should not require external contextual models.

In the rest of this paper, Section 2 describes a feature selection technique, the method of accumulating low and higher levels of co-occurrences and also how interactions are quantified. Section 3 provides experimental details, results and discussion. We conclude in the final section.

2 Method

2.1 Pre-attentive Salient Feature Extraction

Kadir and Brady proposed [3] that saliency is defined as a measure of the unpredictability of a set of data. High unpredictability implies high saliency and vice-versa. Using entropy $\mathcal{H}$, as a measure of statistical unpredictability, high entropy describes data as very salient. Entropy within a local spatio-temporal neighbourhood is defined:

$$\mathcal{H}_D(s_s, s_t, \mathbf{x}) = -\sum_{d \in D} b_{d,s_s,s_t,\mathbf{x}} \log_2 b_{d,s_s,s_t,\mathbf{x}} \quad (1)$$

where s_s is the spatial radius and s_t is the temporal interval of a cylindrical sampling kernel, $\mathbf{x}$ is the point in space and time, around which the cylinder is formed, and d is one of a set of D possible values (e.g. intensity) which are used for approximating the integral of the probability density function as a histogram, b of a local neighbourhood.

Entropy alone is not enough to separate salient from non-salient features. For example, noise would have high entropy since its distribution tends to be quite flat. A more appealing approach measures entropy across multiple scales and attributes a saliency value to a peak in entropy relative to the variation in its intensity distribution at neighbouring scales. This was extended to measure temporal saliency [2], with a two-sided sampling kernel . For this paper, a one-sided version as shown in Figure 1(c), was employed to sample entropy values at different scales from the local spatio-temporal neighbourhood around a grid location $\mathbf{x} = (h, v, t)$. Temporal Saliency, $\mathcal{Y}$ at each location $\mathbf{x}$, is measured as a product of a peak entropy value, $\mathcal{H}$ at its corresponding spatio-temporal scale, and its interscale saliency measure, $\mathcal{W}$.

$$\mathcal{Y}_D(s_{p_s}, s_{p_t}, \mathbf{x}) = \mathcal{H}_D(s_{p_s}, s_{p_t}, \mathbf{x})\ \mathcal{W}_{D_{peak}} \quad (2)$$

where D is a set of possible intensity bins, (s_{p_s}, s_{p_t}) is the spatio-temporal scale at which the entropy, $\mathcal{H}$ peaks, and

$$\mathcal{W}_{D_{peak}} = \mathcal{W}_D(s_{p_s}, s_{p_t}, \mathbf{x})\ \mathcal{W}_D(s_{p_s}, s_{p_t}+1, \mathbf{x}) \quad (3)$$

where $\mathcal{W}_\mathcal{D}$ is a measure of saliency between neighbouring temporal scales,

$$\mathcal{W}_D(s_s, s_t, \mathbf{x}) = s_t \sum_{d \in D} |b_{d,s_s,s_t,\mathbf{x}} - b_{d,s_s,s_{t-1},\mathbf{x}}| \quad (4)$$

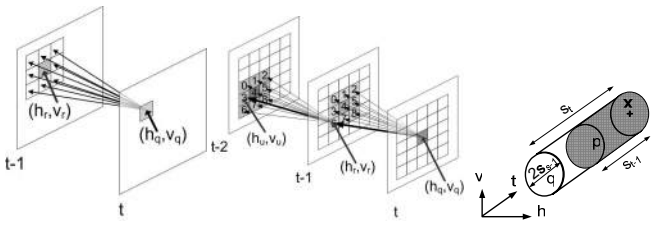
and the spatio-temporal peak is defined as,

$$(s_{p_s}, s_{p_t}) = \{s : s_{s_{peak}} \wedge s_{t_{peak}} \wedge s_{st_{peak}}\} \quad (5)$$

$$\begin{aligned} s_{s_{peak}} &= \mathcal{H}_D(s_s-1, s_t, \mathbf{x}) < \mathcal{H}_D(s_s, s_t, \mathbf{x}) > \mathcal{H}_D(s_s+1, s_t, \mathbf{x}) \\ s_{t_{peak}} &= \mathcal{H}_D(s_s, s_t-1, \mathbf{x}) < \mathcal{H}_D(s_s, s_t, \mathbf{x}) > \mathcal{H}_D(s_s, s_t+1, \mathbf{x}) \\ s_{st_{peak}} &= \mathcal{H}_D(s_s-1, s_t-1, \mathbf{x}) < \mathcal{H}_D(s_s, s_t, \mathbf{x}) > \mathcal{H}_D(s_s+1, s_t+1, \mathbf{x}) \end{aligned} \quad (6)$$

where $s_{s_{peak}}$, $s_{t_{peak}}$, $s_{st_{peak}}$, describes a peak in spatial, temporal, and spatio-temporal scale respectively. Features are selected and quantified with a saliency value and its corresponding spatio-temporal scale. This method is limited to quantifying salient motion at a fixed spatial neighbourhood over time. However, for higher levels of inference, relations need to be drawn between spatially separated but temporally correlated salient motion regions.

2.2 Co-occurrence of Salient Features



(a) (b) (c)

Figure 1: Co-occurrence of saliency values (a) between 2 consecutive frames, (b) using relative orientation based on previous 3 frames. Not all possible co-occurrences have been shown. (c) One sided sampling kernel for calculating entropy variation over spatio-temporal scales.

Establishing co-occurrence of features has been adopted in the past for identifying correlations between spatio-temporally correlated features [10, 8]. Accumulating co-occurrences of features over a sequence generates a well-defined model of likely correspondences. Hence it is possible to quantify how salient the co-occurrence of a local spatio-temporal motion event is based on the inverse of its likelihood. Whilst it may be intuitive to cluster low level features in order to track objects and understand typical scene topology, we argue that this is not practical in crowded scenes where distinguishing between objects is difficult and even fitting well-defined object models are not sufficiently robust [4, 6, 9].

A co-occurrence matrix $\mathbf{N}$ of dimensions $M \times M$ given a set of data $\mathbf{A}$ is defined as the frequency of co-occurrences of all possible combinations of data points, a_q and a_r in $\mathbf{A}$:

$$\mathbf{N} = \sum_{a_q} \sum_{a_r} \Gamma_{ij} \quad (7)$$

where q and r exist between 1 and the cardinality of the set $|\mathbf{A}|$ and the frequency of co-occurrence, Γ_{ij} is

$$\Gamma_{ij} = \begin{cases} 1 & \text{if } a_q \in I_i \wedge a_r \in I_j \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where I_i and I_j define one of M equally distributed intervals or a set of classes, which exist between the maximum and minimum possible values within data set $\mathbf{A}$.

In a more specific case of accumulating the co-occurrence of temporal saliency, the data set $\mathbf{A}$ contains temporal saliency values at every spatio-temporal location $\mathbf{x}$ for the sequence of interest. To reduce complexity, a_q is a temporal saliency value attributed to a particular grid location, (h_q, v_q, t) and a_r define corresponding temporal saliency values within a local spatial neighbourhood in the previous frame. Hence $a_q = \mathcal{H}_D(s_{p_s}, s_{p_t}, h_q, v_q, t)$, $a_r = \mathcal{H}_D(s_{p_s}, s_{p_t}, h_r, v_r, t-1)$, $h_r \in \{h_q - 1, h_q + 1\}$, and $v_r \in \{v_q - 1, v_q + 1)\}$, as shown in Figure 1(a).

If the temporal saliency in two previous frames are used to accumulate co-occurrences, $\mathbf{N}$ is calculated on a frame-by-frame basis.

$$\mathbf{N}(t) = \sum_{a_q} \sum_{a_r} \sum_{u} \Gamma_{ijk} \tag{9}$$

The third dimension takes all possible temporal saliency values a_u from two frames previously so that $a_u = \mathcal{H}_D(s_{p_s}, s_{p_t}, h_u, v_u, t-2)$ where $h_u \in \{h_q - h_r - 1, h_q - h_r + 1\}$, $v_u \in \{v_q - v_r - 1, v_q - v_r + 1\}$. In other words, the 8-pixel neighbourhood is centred around (h_r, v_r). Two extra dimensions are added to $\mathbf{N}(t)$ by introducing co-occurrence of relative orientation of a grid location between consecutive frames as shown in Figure 1(b).

2.3 A Higher Level of Co-occurrence

Higher levels of inference are needed to represent events triggered by spatially separated multiple salient motion events occurring simultaneously (or within a local temporal interval). These levels of activity are detected as unusual multiple simultaneous changes in salient motion, over a local temporal neighbourhood. A two-dimensional co-occurrence matrix, $\mathbf{D}(t)$, accumulated from the difference between $\mathbf{N}(t)$ at two consecutive frames is

$$\mathbf{D}(t) = \sum_{\mathbf{N'}_d} \sum_{\mathbf{N'}_e} \Gamma_{lm} \tag{10}$$

where

$$\Gamma_{lm} = \begin{cases} 1 & \text{if } \mathbf{N'}_d \in I_l \wedge \mathbf{N'}_e \in I_m \\ & \quad \wedge \sum_\tau \beta_{de} \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where $\tau \in \{t-2, t\}$ specifies a local temporal neighbourhood and I_l and I_m exist in equally spaced intervals between the minimum and maximum of $\mathbf{N'} = \frac{d\mathbf{N}(t)}{dt}$ defined as either $\mathbf{N}_d = \mathbf{N}(\tau) - \mathbf{N}(\tau - 1)$ or $\mathbf{N}_e = \mathbf{N}(t) - \mathbf{N}(t-1)$. s_{p_s} define the spatial scale at which the entropy peaked around $\mathbf{x}$. A pair of triples, defined as two elements in matrix $\mathbf{N}(t)$ are evaluated for overlap as:

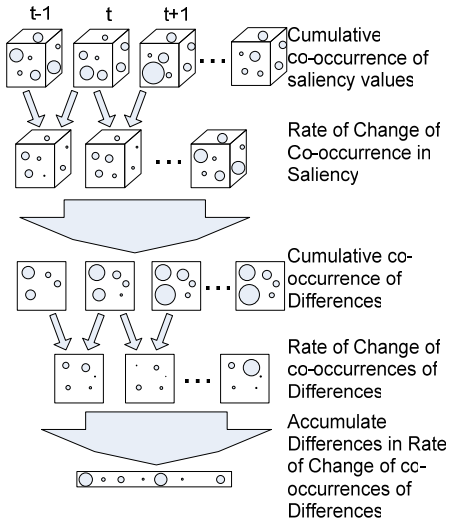


Figure 2: Diagrammatic representation of the matrices used in the co-occurrence algorithm. Cubes in the top row represent the multi-dimensional matrix $\mathbf{N}(t)$ using the co-occurrence method shown in Figure 1(b). The second row of cubes represent the multi-dimensional matrix $\frac{d\mathbf{N}(t)}{dt}$. The next row of squares represent two-dimensional matrices, which are temporal accumulations of the matrix $\mathbf{D}(t)$. Next row shows $\frac{d\mathbf{D}(t)}{dt}$ and the final row shows one-dimensional vector $\mathbf{O}$. Circle diameters represent possible matrix values throughout the progression of the algorithm.

$$\beta_{de} = \begin{cases} 1 & \text{if } (h_d, v_d, t) = (h_e, v_e, t-1) \\ & \quad \wedge s_{p_s}(h_d, v_d, t) = s_{p_s}(h_e, v_e, t-1) \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

Triples are only considered to overlap if two or more of their spatial coordinates and corresponding spatial scales are equal. We assume that a single moving object will not produce multiple peaks in the entropy-scale characteristic. Rather, multiple peaks will be caused by different objects crossing the same spatio-temporal neighbourhood, causing some form of occlusion.

2.4 Quantifying the Interactions

To distinguish between usual and unusual fluctuations in $\mathbf{D}(t)$, we accumulate the frequency of occurrence of $\mathbf{D'} = \frac{d\mathbf{D}(t)}{dt}$ in a one-dimensional vector $\mathbf{O}$ as follows:

$$\mathbf{O} = \sum_{\mathbf{D'}_f} \Omega_o \tag{13}$$

where the occurrence histogram Ω_o is,

$$\Omega_o = \begin{cases} 1 & \text{if } \frac{d\mathbf{D}(t)}{dt}_f \in I_o \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

where $\mathbf{D'}_f = \mathbf{D}(t) - \mathbf{D}(t-1)$ and I_o exists in equally spaced intervals between the minimum and maximum of $\mathbf{D'}$. The least frequent occurrences define the salient events. Hence the inverse of the frequency of occurrence provides a measure of interactive saliency. A diagrammatic representation of the co-occurrence algorithm is shown in Figure 2.

3 Experiment

Experiments were carried out on two contrasting sequences of a busy traffic scene and corridor entrance scenario. Typical frames of the two scenes are shown in Figure 3.

Figure 3: Typical frames from the two scenes. Top: busy traffic. Middle: corridor entrance.

3.1 Basic Co-occurrence

The co-occurrence of saliency features based on Figure 1(a) was calculated for a busy traffic scene containing 3100 frames sampled at 25Hz, and subsampled 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(a) where a selection of the frames highlighting the 10% least frequent co-occurrences of matrix $\mathbf{N}(t)$ are shown. The leftmost frame of Figure 4(a) identifies a car slowing down and turning to change lanes. The rest of row (a) shows a detected instance of a reversing car.

Figure 4: Frames illustrating the ranked top 10% least frequent co-occurrence elements from a busy traffic scene (row(a)), and a corridor and entrance scene (rows (b,c)) using the basic method from Eqn. (7). 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.

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. Figure 5 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. 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 stopped reversing and changed lanes (see top left corner in Figure 5), this was not considered by the algorithm to be globally salient. This suggests that a hierarchical co-occurrence model should be employed. 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 spatio-temporal volumes would remove sensitivity to perspective variations. This is demonstrated in our results using the higher-level hierarchical model later in Section 3.2.

Figure 5: 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.

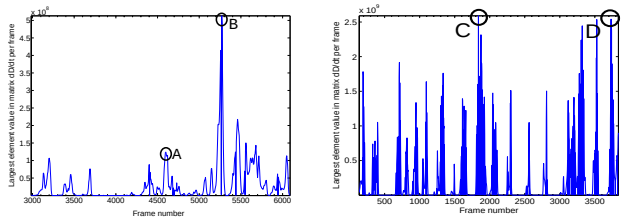
Similar results are shown for a scene of a secure entrance in a corridor, containing quite complex motion patterns to three possible entrances/exits. The sequence consisted of 4000 frames taken at 10Hz, and sub-sampled by 3 frames. Figure 4(b,c) shows the spatio-temporal location of the least frequent 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(c) shows a person running to catch the door, and also two 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). To detect salient directions of motion and reduce the number of false-positives, the co-occurrence method of Eqn. (7) was extended to calculate the relative orientation of co-occurring saliency values. 10 bins were used to accumulate the saliency values for three dimensions of the co-occurrence matrix with two extra dimensions for the number of orientations between neighbouring grids in the previous frames, consisting of 9 bins for each pair of consecutive frames. Despite the additional orientation and data from

two previous frames, the results were very similar to the basic method. Therefore higher levels of co-occurrence are needed.

3.2 Temporal Correlation and Quantification

The method, as described in Figure 2 was employed, using 20 bins to accumulate correlations of the rate of change of co-occurrence of saliency values $\mathbf{N}'$. Rather than assuming temporal correlations occurred simultaneously, a temporal interval of three subsampled consecutive frames was used. For complexity reasons, the algorithm was streamlined by not taking into account all the normal fluctuations to matrix $\mathbf{D}(t)$. The ratio of usual-unusual fluctuations were high enough that the results would not be affected by ignoring some normal data. Figure 6 shows the maximum of matrix $\mathbf{D}'$ at each frame, as described in Section 2.3, over both sequences. Peaks show salient spatially separated but temporally correlated motion.



(a) (b)

Figure 6: Plot showing salient temporally correlated motion events in each frame for scenes of (a) busy traffic (b) corridor entrance. Peaks highlighted, show the frames containing the two reversing car events. Peak A:first (undetected) reversing car event. Peak B:detected second reversing car event. Peak C: person running to catch the door. Peak D: two people who can't open the door and turn back.

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 7. The second reversing car event is detected very clearly in Figure 7(g-i) where salient correlations were found between the reversing car (that was slowing down before reversing) and other normally behaving cars within the scene. The first (previously undetected) reversing car is still not detected but a car that has to slow down and change lanes due to the reversing car, is detected. Salient correlations were found between the affected car and surrounding cars that were behaving normally (see Figure 7(d-f)). Some of the peaks in Figure 6(a) correspond to the two reversing car events.

Figure 7(a-b) highlights the event of the middle car 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. Unfortunately, the car on the left is also correlated with its itself three frames previously, though this can be easily eliminated by adjusting the time interval over which temporal correlations are accumulated. Figure 7(c,j) highlighted problems caused by the assumption that no triples should overlap if the corresponding spatial scales at which the entropy peaked were the same. The thickness of the lines between interacting objects provides a clearly distinguishable difference between more and less salient interactions.

The results from the corridor scene also showed similar separation between frequencies of occurrence of matrix $\mathbf{O}$, all the infrequent occurrences were at least 2.6×10^{-05} times smaller than the most frequently accumulated bin. Compared to the busy traffic scenario, 18 out of the 20 bins were filled. Since the scene was much less busy compared to the traffic sequence, all events involving opening the secure doors were detected as salient, as shown by the large numbers of peaks in Figure 6(b).

The two salient motion events detected in the previous section were also detected using this method, as shown in Figure 7(p-t), which were amongst the least frequent occurrences. The first salient event where someone rushed to catch the door, is shown in (p,q) where correlations were confined to the area at the top of the frame, rather than between running person and the closing door. The second correctly detected salient event was the two people who can't get in and turn round in (r-t). The correlated activity between the heads of the two people was detected, as shown in (t). Note that the interaction between the heads was considered less salient than loitering round the door for an unusual amount of time. Many correlations were made between the person opening the door and the motion of the door, such as those shown in (l,n,o). Again, there were some correlations of the motion of a person with themselves in a later frames, as shown in (k-t). In (k,m), correlations were also made between the person and their reflection in the glass of the secure doors. Discrimination between salient activities was apparent, though better discrimination between different interactions would be needed to facilitate a more complex and informative hierarchical structure.

4 Conclusion

We have demonstrated that, it is possible to perform high-level inferences about scene dynamics using a bottom-up approach through lower and higher levels of co-occurrence. More importantly, we have shown that significant high-level dynamics of a scene can be detected from the scene data alone. At the start of the paper, we 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, we have shown that it is possible to use the same numbers of bins for two contrasting sequences where the depth of perspective and also size of the moving objects, and typical object trajectories were very differ-

Figure 7: Frames showing the least frequent occurrence elements from **O** from a busy traffic (a-h), and a corridor scene (i-p). 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.

ent. A weakness of the algorithm is that detection of some unusual salient motion relies on usual behaviour to occur within the local temporal neighbourhood. However, salient motion events that affect the typical behaviour patterns in other parts of a scene would be considered more salient than a lone object behaving abnormally. We also demonstrate an effective salient interaction quantification measure. This could be made more discriminative if many of the overlapping triples could be eliminated.

Further work will be carried out to better monitor the evolution of the co-occurrence matrix and hence detect salient interactions more efficiently. The algorithm also needs to be cleaned up so clusters of triples can be treated as motion from the same object. This should facilitate better quantification and lead to more complex hierarchical modelling of the scene data for classification purposes. Currently, the method works completely offline so it follows that an on-line version would be beneficial.

References

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

[2] Hayley Hung and Shaogang Gong. Quantifying temporal saliency. In *BMVC*, pages 727–736, September 2004.

[3] T Kadir and M Brady. Scale, saliency and image description. *IJCV*, 45(2):83–105, November 2001.

[4] Zia Khan, Tucker Balch, and Frank Dellaert. An mcmc-based particle filter for tracking mutiple interacting targets. *GIT Technical Report*, GIT-GVU-03-35, October 2003.

[5] Tony Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30(2):77–116, 1998.

[6] Songhwai Oh, Stuart Russell, and Shankar Sastry. Markov chain monte carlo data association for general multiple-traget tracking problems. In *IEEE Conference on Decision and Control*, 2004.

[7] B Schiele and J Crowley. Object recognition using multidimensional receptive field histograms. In *ECCV (1)*, pages 610–619, 1996.

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

[9] Tao Zhao and Ram Nevatia. Tracking multiple humans in crowded environment. In *CVPR*, 2003.

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