Quantifying Temporal Saliency

Hayley Hung and Shaogang Gong
Queen Mary Vision Laboratory
Queen Mary University of London
E1 4NS
{hswh,sgg}@dcs.qmul.ac.uk

Abstract

A significant problem in automatic scene interpretation is the ability to perform contextually meaningful segmentation of both static and moving images using a bottom-up approach. We examine and propose an extension to Kadir and Brady's Scale Saliency Algorithm for quantifying temporal saliency and performing automatic spatial and temporal scale selection.

1 Introduction

Accurate interpretation of a visual scene is a highly selective process that requires us to ignore some aspects of the visual information, deemed to be either noise or irrelevant, depending on the context. The effectiveness of this selection process depends heavily on how context is represented. Intuitively, contextual information about a visual scene lends itself naturally to a top-down based representation. However, a significant problem in such a representation of contextual knowledge is that it is not readily measurable by the information that can be extracted from the imagery data without being made too scene-specific. It is therefore both necessary and attractive to develop a bottom-up model that is capable of extracting image features which represent something contextually meaningful. We consider a key to solving this problem is the ability to quantify holistically the spatial and temporal saliency of local image features.

Popular techniques for extracting salient local imagery features adopt orientation filters to identify what is considered to be a salient part of an image [5, 1]. However, we argue that such information is not neccessarily salient since features that produce a higher magnitude reponse from orientational filters are largely dependent on the choice of the

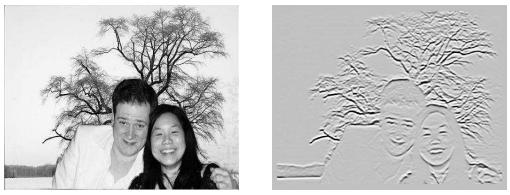


Figure 1: The picture on the right shows the effects of just applying a horizontal first order gaussian filter to the image on the left. Although the people in the foreground are clearly of more interest in the picture, it is the cluttered background that exhibits the highest magnitude responses.

basis functions which can be rather arbitrary. If an object of interest is placed against a cluttered background, using orientational filter responses to identify salient parts in an image would mean that even the cluttered background could be extracted readily as a significant part of the scene. The problem is clearly shown by the example in Figure 1.

A statistical approach to image description might be more sensible since using a probabilistic representation means that no information is thrown away by taking local considerations alone. An example of this approach was proposed by Kadir and Brady [3] in which entropy is used 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. They argue that saliency is local unpredictability or high entropy, where local entropy is defined as:

$$\mathcal{H}_D(s,\mathbf{x}) = -\sum_{d\in D} p_{d,s,\mathbf{x}} \log_2 \left(p_{d,s,\mathbf{x}}\right) \tag{1}$$

where $p_{d,s,\mathbf{x}}$ is the probability density function (PDF) of a local neighbourhood prescribed by the scale or radius of this neighbourhood s, d is one of a set of D possible values which are used for approximating the integral of the PDF as a histogram (e.g. intensity values), and $\mathbf{x}$ is the point around which the local neighbourhood is defined. The novelty of this approach is that features are selected based on the variation of the entropy ($\mathcal{H}_D$) over different scales. The entropy-scale characteristics of a particular local neighbourhood represents the local image structure or local *context*. Kadir and Brady proposed using an inter-scale saliency measure $\mathcal{W}_D$ to describe the relation between entropy of a local neighbourhood at two different scales:

$$\mathcal{W}_D(s,\mathbf{x}) = \frac{s^2}{2s-1} \sum_{d\in D} \left|p_{d,s,\mathbf{x}} - p_{d,(s-1),\mathbf{x}}\right| \tag{2}$$

The scale at which the entropy peaks is deemed the most suitable scale to describe an image since it is the scale at which the image becomes unpredictable or more difficult to model [4]. Kadir and Brady further suggested that a saliency measure can be obtained from the scalar product of the entropy and inter-scale saliency at the scale value s_p at which the entropy peaks:

$$\mathcal{Y}_D(s_p,\mathbf{x}) = \mathcal{H}_D(s_p,\mathbf{x})\ \mathcal{W}_D(s_p,\mathbf{x}) \tag{3}$$

This was designed to take into account to some extent the local nature of the peak. The scale at which the peak entropy occurs was defined as:

$$s_p = \{s : \mathcal{H}_D(s-1,\mathbf{x}) < \mathcal{H}_D(s,\mathbf{x}) > \mathcal{H}_D(s+1,\mathbf{x})\} \tag{4}$$

However, Kadir and Brady's model of saliency using an entropy measure between two adjacent scales does not always depict saliency accurately. Figure 2 shows an example of the variation of entropy over scale for three image regions of different characteristics. The conclusion is that since the eye region has a much higher entropy value and a flatter pdf, it exhibits much more unpredictable behaviour than the sky region. As expected, we have been able to separate foreground from background. However, the entropy values of the cluttered region of the tree are the largest and exhibits highly unpredictable behaviour, though we expect this region to be background. But inspection of the entropy-scale characteristic shows that the tree region has a more similar response to the sky region. So the entropy-scale curve tells us much more about the inherent regions of interest in the image.

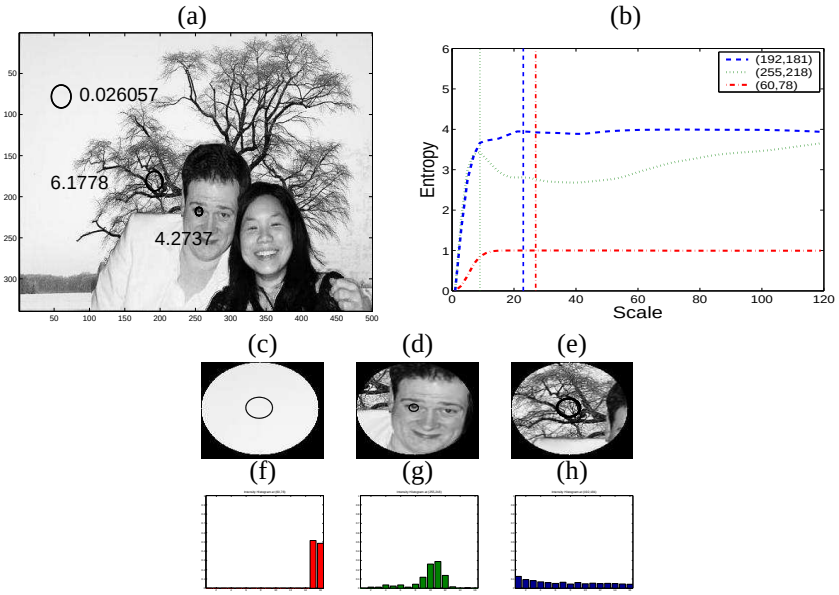


Figure 2: (a) A cluttered image with, saliency values calculated at three contrasting points using the spatial saliency algorithm.(b) The entropy-scale characteristics of the three regions, where the vertical lines indicate the lowest scale that corresponds to a peak in entropy. (c-e) Enlarged versions 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. The PDF for the eye region exhibits are much flatter compared to that of the sky region. Hence their corresponding entropy-scale characteristics show that the eye region has a much higher entropy value at all scales. However, the cluttered background of the tree region has a much flatter PDF and a higher entropy value than either the eye or sky regions. The entropy-scale characteristic indicates in (b) that the sky and tree region exhibit very similar variations compared to that of the eye region.

Referring to Figure 2 (a) the eye region has a more pronounced peak in entropy and hence a higher saliency score compared to the sky region. However, the cluttered background region has an even higher saliency value than either of the other two regions. This suggests that calculating the inter-scale measure $\mathcal{W}_D$ from just one adjacent scale is not enough.

Automatic scene interpretation of a 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 advantages of extracting features using an entropy-scale measure is that it can represent a statistical model of the variation of a particular image region over space and time, potentially capable of separating foreground from non-stationary background regions in a scene, and identifying temporally salient patterns of change.

In this paper, we extend Kadir and Brady's scale saliency model to quantifying temporal saliency for performing automatic spatial and temporal scale selection. In Section 2, we introduce the concept for bridging the gap between spatial saliency and temporal

saliency. In Section 3 we formulate the notion of temporal saliency. A brief description of the algorithm is provided in Section 4. Experiments are shown in Section 5 and we conclude in Section 6.

2 From Spatial to Temporal Saliency

Let us first illustrate the potential in taking into account more than just two adjacent scales in measuring saliency. We extend the scale saliency equation to inter-scale saliency on both side of a peak as follows:

$$\mathcal{Y}_D(s_p, \mathbf{x}) = \mathcal{H}_D(s_p, \mathbf{x})\ \mathcal{W}_{D_{peak}} \tag{5}$$

where

$$\mathcal{W}_{D_{peak}} = \mathcal{W}_D(s_p, \mathbf{x})\ \mathcal{W}_D(s_p + 1, \mathbf{x}) \tag{6}$$

The new term $\mathcal{W}_{D_{peak}}$ measures the inter-scale entropy between the peak scale and the next scale up. The immediate effect of the new interscale measure can be seen using the previous examples shown in Figure 2. The saliency measure for the cluttered, the eye and the sky regions have changed from $(6.18, 4.27, 0.03)$ to $(6.94, 8.05, 0.00)$ respectively. We further suggest that this inter-scale measure becomes more significant if the regions of interest in the image are non-stationary over time.

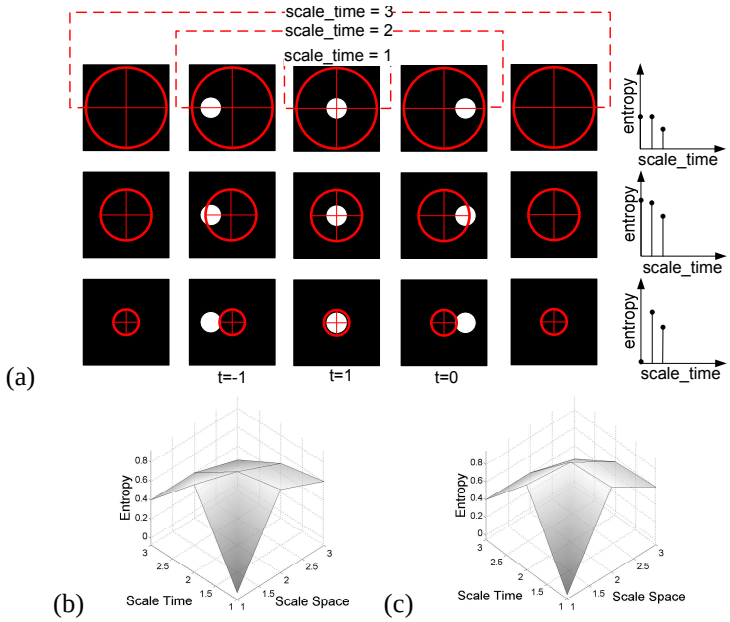
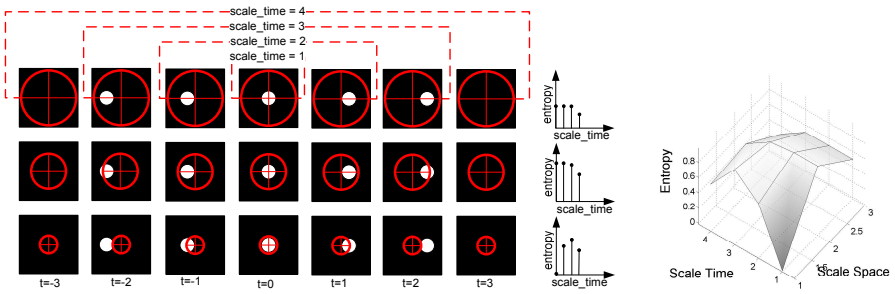


Figure 3: Entropy-Scale characteristics of a dot passing over a plain background. (a) the same sequence of images with three different sizes of kernel. (b) the entropy scale characteristic centred at $t = -1$. (c) the entropy-scale characteristic at $t = 0$.

To this end, we extend the notion of saliency from a measure of spatial unpredictability to temporal unpredictability. At its most simple form, something temporally unpredictable occurs when a particular intensity distribution appears or disappears from a sampled neighbourhood. Figure 3 shows a synthesised sequence of a white dot moving from



(a) (b)

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

left to right on a black background. The manually calculated graphs in the rightmost column show 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 demonstrates 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 4, then the peak in entropy is seen at a higher time scale.

3 Quantifying Temporal Saliency

To calculate temporal saliency, we consider that PDFs at different scales are generated from spatiotemporal cylinders where the entropy at a particular spatiotemporal cylinder is defined as:

$$\mathcal{H}_D(s_s, s_t, \mathbf{x}) = -\sum_{d \in D} p_{d,s_s,s_t,\mathbf{x}} \log_2 p_{d,s_s,s_t,\mathbf{x}} \tag{7}$$

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

The inter-scale saliency measure $\mathcal{W}_D$ becomes a two dimensional matrix, representing

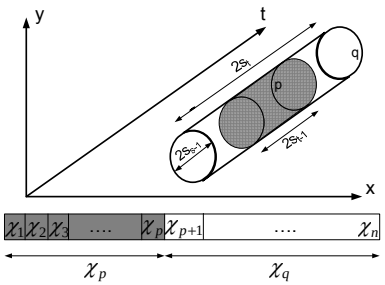


Figure 5: Inter-scale volumes.

the differential of the entropy for varying scales in time and space. A normalisation factor is calculated for cylinderical volumes rather than circles [2]. As shown in Figure 5, the set $\chi_q \cup \chi_p$ is used to create the PDF at scale $\{s_s, s_t\}$ and χ_p is used to create the PDF at scale $\{s_s, s_t - 1\}$. Hence intertemporal scale saliency is defined as:

$$\mathcal{W}_D(s_s, s_t, \mathbf{x}) = s_t \sum_{d \in D} |p_{d,s_s,s_t,\mathbf{x}} - p_{d,s_s,s_t-1,\mathbf{x}}| \tag{8}$$

where s_t is the inter-scale normalisation factor for the cylinderical volumes shown in Figure 5. Taking into account both sides of the peak, $\mathcal{Y}_\mathcal{D}$ is again defined as the scalar product of the inter-scale saliency measure and its entropy value evaluated at $\{s_{p_s}, s_{p_t}\}$:

$$\mathcal{Y}_D(s_{p_s}, s_{p_t}, \mathbf{x}) = \mathcal{H}_D(s_{p_s}, s_{p_t}, \mathbf{x})\ \mathcal{W}_{D_{peak}} \tag{9}$$

where

$$\mathcal{W}_{D_{peak}} = \mathcal{W}_D(s_{p_s}, s_{p_t}, \mathbf{x})\ \mathcal{W}_D(s_{p_s}, s_{p_t} + 1, \mathbf{x}) \tag{10}$$

where

$$s_p = \{s : s_{s_{peak}} \wedge s_{t_{peak}} \wedge s_{st_{peak}}\} \tag{11}$$

and

$$\begin{array}{rcccccc} s_{s_{peak}} & = & \mathcal{H}_D(s_s-1, s_t, \mathbf{x}) & < & \mathcal{H}_D(s_s, s_t, \mathbf{x}) & > & \mathcal{H}_D(s_s+1, s_t, \mathbf{x}) \\ s_{t_{peak}} & = & \mathcal{H}_D(s_s, s_t-1, \mathbf{x}) & < & \mathcal{H}_D(s_s, s_t, \mathbf{x}) & > & \mathcal{H}_D(s_s, s_t+1, \mathbf{x}) \\ s_{st_{peak}} & = & \mathcal{H}_D(s_s-1, s_t-1, \mathbf{x}) & < & \mathcal{H}_D(s_s, s_t, \mathbf{x}) & > & \mathcal{H}_D(s_s+1, s_t+1, \mathbf{x}) \end{array} \tag{12}$$

where $s_{s_{peak}}$ is a spatial scale peak, $s_{t_{peak}}$ is a temporal scale peak, and $s_{st_{peak}}$ is a spatiotemporal scale peak. It is important to clarify that this definition of temporal saliency favours peaks found in both inter-spatiotemporal scale-space with high inter-temporal scale differences in PDF, whilst only inter-temporal scales are used to calculate $\mathcal{W}_{D_{peak}}$. It is likely that if a inter-spatiotemporal saliency measure was used, the inter-spatial element would need to have a smaller weighting on the measure to avoid sensitivity to cluttered background features.

4 The Temporal Saliency Algorithm

1. Firstly, the image is divided into a grid for reduced computation. This defines the granularity of the motion and size of objects to be detected in the image. If there is a high degree of variation in entropy within a grid, the algorithm is performed on a subdivided part of that grid.
2. The entropy-scale characteristics, $\mathcal{H}_D$ and inter-scale saliency, $\mathcal{W}_D$ is calculated, centred at each grid location and time frame for all $\{\{s_s, s_t\} : \{1,1\} \leq \{s_s, s_t\} \leq \{s_{s_{max}}, s_{t_{max}}\}\}$.
3. A saliency measure is calculated based on the entropy value between temporal scales at the location of the peaks in entropy over temporal, spatial and spatiotemporal scales.
4. After a suitable number of frames have been sampled, the peaks in entropy are ranked according to their corresponding saliency value.
5. The most salient time intervals and spatial regions are located.

5 Experiment

Experiments were run for 4 different indoor and outdoor sequences. The first is a sequence from an outdoor scene that is highly cluttered and under constant change. The pitfalls of analysing this scene is that there are rapid variations in lighting and the windy conditions means that a bush at the bottom of each frame, is prone to a lot of motion. The wind also causes the camera to move on several occasions, causing the entire scene to shift slightly. Each frame was divided into a 7×10 grid. For each grid location, the entropy-scale characteristic was calculated. Saliency values were calculated for any spatiotemporal peaks that were found where $\{s_{s_{max}}, s_{t_{max}}\}$ was $\{20, 70\}$, and the pixel grid size was 31.

The ranked saliency results indicated two clearly separable events as shown in frames $\{$F-H$\}$ and $\{$I-K$\}$ of Figure 6. The top row of Figure 6 shows three different methods for locating salient behaviour. It is clear to see that the temporal saliency algorthm does not pick out as much global motion as temporal frame differencing and hence there is some background motion supression. In particular, the noisy motion of the bush at the bottom of the scene had much lower saliency values, compared to that of the moving car. The highest saliency value caused by the bush and moving car was 0.55 and 8.37 respectively. The strong background motion supression is further demonstrated by Figure 6 (e) where the saliency is plotted in order of rank for a typical frame of the sequence. In Figure 6 (b), a higher percentage of salient locations are shown for the spatial saliency algorithm since this produced much fewer salient results.

Figure 7 shows 3 of the most salient frames for 3 different sequences from both indoor and outdoor scenes. In each case, meaningful salient motion, within the context of the chosen sequnces, was detected. Column (a) of Figure 7 shows the results of from an extremely busy traffic scene. Within this sequence, there are are many cars that move along the roads. However, at a certain point, a car stops and reverses onto oncoming traffic. The grid size was set particularly small at 11 in just this set of results in order to accomodate for the large amounts of diagonal motion in the scene. There is 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 that was detected and is highlighted at peak B of Figure 6. However, the saliency value of this particular event is not the highest for the saliency measures calculated, centered at this frame.

Column (b) of Figure 7 shows the results from a simulated scene of a drinks shop. The scene is highly cluttered and the shop keeper to the right of the frame causes some noisy background motion but this 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 column (b) indicates time-localised clusters of salient motion.

Column (c) of Figure 7 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 experiment was run over a relatively short sequence so 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.

6 Conclusion

We have shown that it is possible to quantify effectively, different types of temporally varying activity using a purely bottom up approach. The problems with automatically interpreting real scenes is that it is very difficult to separate noisy background clutter from salient motion. Hence using a contextually meaningful approach has meant that whilst the noisy background clutter is detected, its response tends to be much lower in comparison to other more salient parts of a scene. We have demonstrated that it is possible to suppress such background clutter, as well as small motions from a moving camera, and subtle lighting changes, using a relatively simple algorithm.

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. Due to the algorithm's sensitivity to shift, it is possible to some extent, to track salient objects in a scene, which is shown most clearly in column (c) of Figure 7. A simple modification would be to correlate the salient locations with the resultant regions found using consecutive frame temporal differencing. However, this would not provide information about temporally salient intervals.

The algorithm needs to be modified so that it takes into account inter-spatial and inter-temporal entropy-scale responses. Hence some form of meaningful tracking can be achieved from clustering based on, for example, the entropy scale characterstic between a local neighbourhood of grid cells. Furthermore, it will be possible to reduce computational complexity by using dynamic programming to concentrate computation on more localised areas that are deemed to be more salient. Another method to reduce computation, would be to group salient activities into different levels of levels of complexity using a hierarchical approach. This would also aid classification the process.

As yet, the algorithm is not able to distinguish between rareity and saliency. In one particular experiment, the most salient feature of a busy traffic scene was the light reflecting off a vehicle. Such anomalies would be ranked out of significance if the algorithm was run over a longer time intervals, but only provided enough occurrences were found. However, if the algorithm could generate a model of the scene whilst detecting salient behaviour, it will be able to incorporate this into its computation and further suppress noisy background clutter.

Overall, the algorithm shows great potential for a bottom-up approach to scene interpretation. It is able to supress background clutter, though modifications need to be made in order to segment salient spatiotemporal regions within a scene.

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

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

[4] T Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30(2), 1998.

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

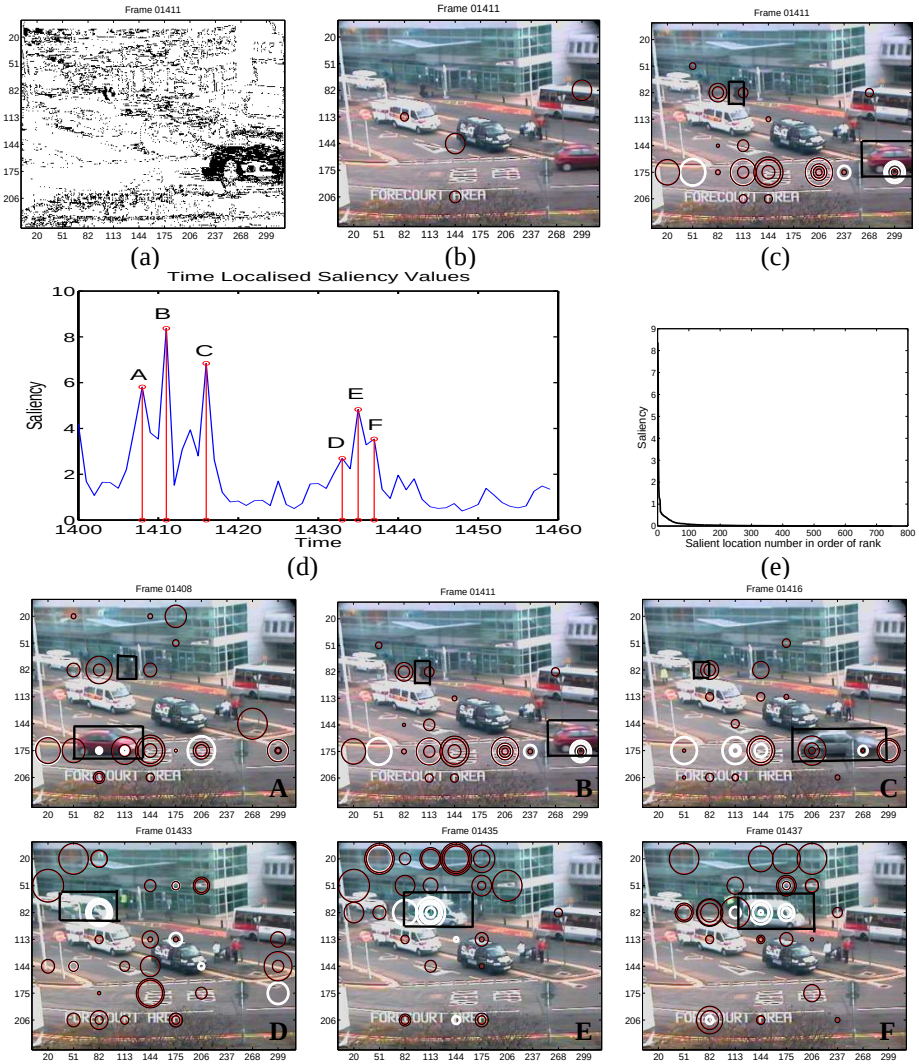


Figure 6: Detailed study of the results from a highly cluttered scene run over 60 frames. (a) consecutive frame temporal differencing using a threshold of 20, (b) 10% most salient regions using the spatial saliency algorithm, and (c) the 5% most salient regions using the temporal saliency algorithm. (d) The highest saliency value at each frame. (e) Typical saliency values centred at one frame plotted against rank. Bottom two rows show the central frames indicated in (d) for peaks (A-F) respectively. The circles in show the spatial location of the salient motion: black circles are the top 10% in (b) and top 5% in the rest, white circles are the 1% most salient locations, black boxes are manual indications of where salient regions exist. The size of the circles indicates the spatial scale at which the entropy peaked. If the circles do not appear in a visually salient region, it is likely that salient motion occurred before or after the current frame. The circles appear symmetrically in time about the central frame hence it is not possible identify whether salient motion at a particular time scale occured before or after this central frame. For clarity, only salient regions whose temporal peak occured within a sensible interval around the central frame are shown.

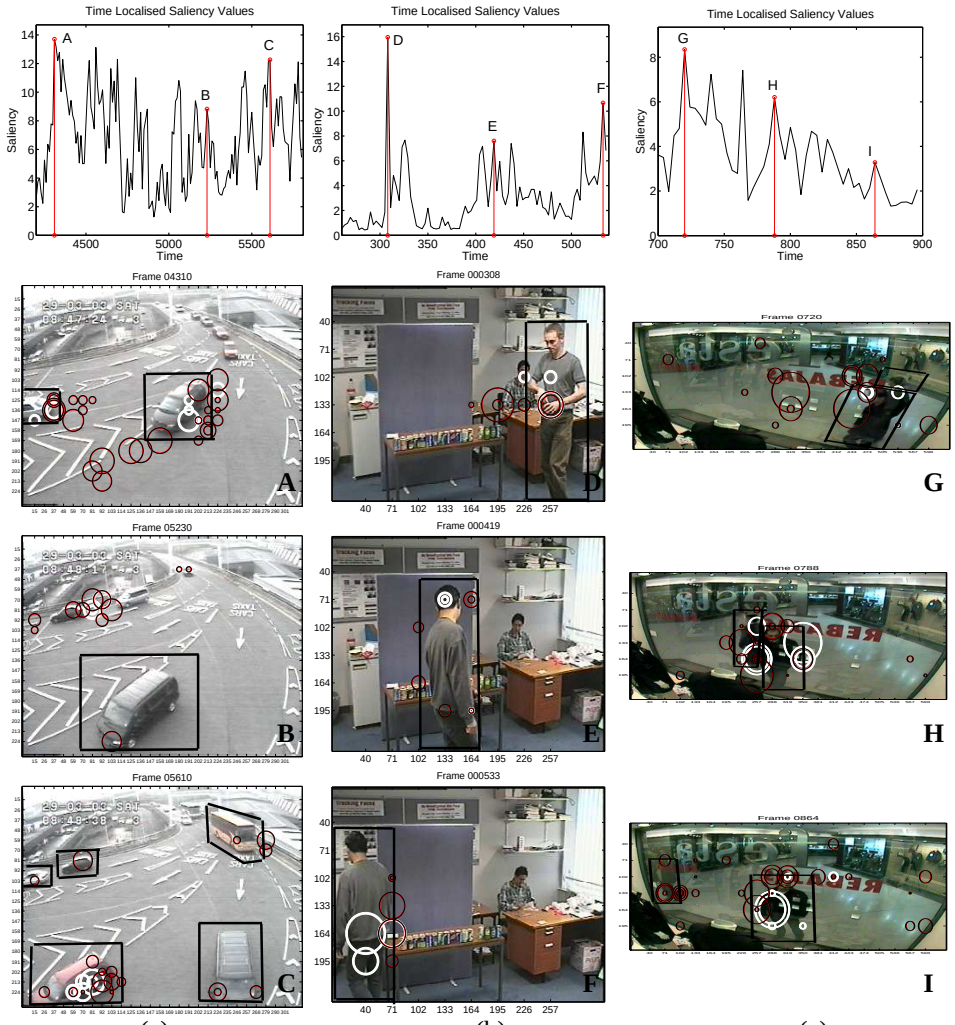


Figure 7: Results from a mixture of different indoor and outdoor scenes. (a) Outdoor scene of a busy road run over 1400 frames. (b) Simulation of a drinks shop run over 372 frames . (c) Cluttered scene through a shop window run over 150 frames. The circles indicate regions of salient motion. The top row shows the highest saliency value centered 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 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 .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. Clearly, from inspection of the frames, it is possible to interpolate the salient motion of the objects.