# A SALIENCY BASED OBJECT TRACKING METHOD

*Shijie Zhang and Fred Stentiford*

University College London, Adastral Park Campus, Ross Building
Martlesham Heath, Ipswich, IP5 3RE, UK
{j.zhang, f.stentiford}@adastral.ucl.ac.uk

## ABSTRACT

A novel three-stage framework for object tracking under stationary background conditions is proposed in this paper. The first stage uses an attention based method to extract motion information. The second stage then applies a region growing and matching technique to motion vectors to obtain motion segmentation. Finally the moving objects are tracked based on the displacements of region centroids. The method is tested on various real-world video data and empirical results show that the proposed approach can track moving objects and extract motion information from non-rigid objects such as moving people without prior knowledge of the object's size or shape.

## 1. INTRODUCTION

The demand for automated motion detection and object tracking systems has promoted considerable research activity in the field of computer vision [1-9]. This paper proposes a method to detect and measure motion based upon tracking salient features using a model of visual attention.

Stauffer and Grimson [1] presented a novel probabilistic method for background subtraction for multiple object tracking. It modeled each pixel as a separate mixture Gaussian model. After the background subtraction process, foreground pixels were identified, labeled and grouped into regions by a connected components algorithm. The model was updated with an on-line approximation. It copes well with lighting changes, repetitive motions from clutter, and long-term scene changes with different weather conditions involving different cameras. However, problems arise when moving objects occlude each other and one object enters the scene while another is leaving. In addition moving shadows are not removed during tracking. Comaniciu [2] proposed a method for real time non-rigid object tracking with a moving camera based on the mean shift algorithm. One advantage is that the intense blurring due to camera motion did not affect the tracker performance, which is always a problem for contour based trackers. The tracker handles partial occlusions, background clutter and target scale variations. It also works under low quality sequences with compression artifacts but the coarse appearance models can fail to track accurately regions that share similar statistics (colour) with nearby regions. Shi and Tomasi [3] proposed a method for feature selection, a tracking algorithm based on affine change models, and a technique for monitoring features during tracking. The feature selection criterion depended entirely on how well the tracker worked. The tracking algorithm extended the previous Newton-Raphson style search methods to work under affine transformations. The bad features were abandoned based on a measure of dissimilarity that used an affine motion model. The proposed method works under occlusions. However, the method cannot handle deformable objects and the iterative tracking algorithm can take a relatively long time to converge. In [4] a new approach was proposed for visual tracking using dynamic geodesic snakes. The method combined state information (velocity) with every particle on a contour described by a level set function. It works under partial occlusions but also tracks shadows due to the edge-based contour model. Isard and MacCormick [5] proposed a multiple-person tracking system for single camera real-time surveillance applications. A multi-blob likelihood function was adapted from the theory of Bayesian correlation based on learned statistics, but assumed a static camera to create more specific background and foreground modeling. Then a Bayesian multiple-object particle filter was used for tracking. The observation model proposed for object likelihood used a synthesis of learnt background patches, pooled foreground patches and geometric reasoning from the camera calibration. It works with background clutter. One disadvantage of this approach is its failure under occlusions. Bouthemy [6] proposed a novel probabilistic parameter-free method for detecting independently moving objects using the Helmholz principle. Optical flow fields were estimated without making assumptions on motion presence and allowed for possible illumination changes. The method imposes a requirement on the minimum size for the detected region and detection errors arise with small and low contrast objects. Black and Jepson [7] proposed a method for optical flow estimation based on the motion of planar regions plus local deformations. The approach used brightness information for motion interpretation by using segmented regions of piecewise smooth brightness to hypothesize planar regions in the scene. The proposed method has problems dealing with small and fast moving objects. It is also computational expensive. Black and Anandan [8] then proposed a framework based on robust estimation that addressed violations of both brightness constancy and spatial smoothness assumptions caused by multiple motions. It was applied to two common techniques for optical flow estimation: the area-based regression method and the gradient-based method. To cope with motions larger than a single pixel, a coarse-to-fine strategy was employed in which a pyramid of spatially filtered and sub-sampled images was constructed. Separate motions were recovered using estimated affine motions, however, the method is relatively slow. Viola and Jones [9] presented a pedestrian detection system that integrated both image intensity (appearance) and motion information, which was the first approach that combined motion and appearance in a single model. The system works relatively fast and operates on low resolution images under difficult conditions such as rain and snow, but it does not detect occluded or partial human figures.

The use of visual attention (VA) methods [10-13] to define the foreground and background information in a static image for scene analysis has motivated this investigation. We propose in this paper that similar mechanisms may be applied to the detection of saliency in motion and thereby derive an estimate for that motion. The object tracking framework is presented in Section 2. Results are shown in Section 3 along with some discussion. Finally, Section 4 outlines conclusions and future work.

## 2. OBJECT TRACKING FRAMEWORK OUTLINE

The proposed framework contains three stages. The first stage uses an attention based method to estimate and extract motion information [14]. The second stage then applies a region growing and matching technique to motion vectors extracted [15]. In the last stage moving objects are tracked based on linking region centroids. The outline of the method is given below.

### 2.1. Motion estimation based on visual attention

Regions of static saliency have been identified using an attention method described in [12]. Those regions which are largely different to other parts of the image will be salient and are likely to be in the foreground. This concept has been extended into the time domain and is applied to frames from video sequences to detect salient motion. The approach [14] does not require an initial segmentation process and depends only upon the detection of anomalous movements. The method estimates the shift of locations between frames by obtaining the distribution of displacements of corresponding salient features around these locations.

In this paper candidate regions of motion are detected by generating the intensity difference between the current frame and a background reference frame obtained by averaging a series of frames in an unchanging video sequence. A threshold is then applied producing a *potential motion template.* The intensity difference $I_x$ between pixels *x* in the current frame and the reference is given by

$$I_{\boldsymbol{x}} = \{|r_2 - r_1| + |g_2 - g_1| + |b_2 - b_1|\}, \tag{1}$$

where parameters $(r_1, g_1, b_1)$ & $(r_2, g_2, b_2)$ represent the rgb colour values for pixel $\boldsymbol{x}$ in reference frame and the current frame. The intensity $I_x$ is calculated by taking the sum of the differences of rgb values between the two frames.

The candidate regions $R_t$ in the frame t are then identified where $I_x > T$ where $T$ is a fixed threshold

Let a pixel $\boldsymbol{x} = (x, y)$ in $R_t$ correspond to colour components $\boldsymbol{a} = (r, g, b)$. Let $\boldsymbol{F(x)} = \boldsymbol{a}$ and let $\boldsymbol{x_0}$ be in $R_t$ in frame t. Consider a neighbourhood $G$ of $\boldsymbol{x_0}$ within a window of radius $\varepsilon$ where

$$\{\boldsymbol{x}'_i \in G \quad iff \quad |x_0 - x'| \le \varepsilon\}. \tag{2}$$

Select a set of $m$ points $S_x$ in $G$ (called a fork) where

$$S_x = \{\boldsymbol{x}'_1, \boldsymbol{x}'_2, ..., \boldsymbol{x}'_m\}. \tag{3}$$

Forks are only generated which contain pixels that mismatch each other. This means that they are selected in image regions possessing high or certainly non-zero attention scores, such as on edges or other salient features as reported earlier [12].

In this case the criteria is set so that at least one pixel in the fork will differ with one or more of the other fork pixels by more than $\delta$ in one or more of its rgb values i.e.

$$\left|F_k(\boldsymbol{x}'_i) - F_k(\boldsymbol{x}'_j)\right| > \delta_k, \quad for\ some\ i, j, k. \tag{4}$$

Define the radius of the region within which fork comparisons will be made as $V$ (*view radius*). Select another location $\boldsymbol{y_0}$ in the next frame region $R_{t+1}$ within a radius $V$ of $\boldsymbol{x_0}$.

Define the second fork

$$S_y = \{\boldsymbol{y}'_1, \boldsymbol{y}'_2, ..., \boldsymbol{y}'_m\} \text{ where } \boldsymbol{x_0} - \boldsymbol{x}'_i = \boldsymbol{y_0} - \boldsymbol{y}'_i \tag{5}$$
$$\text{and} \left|\boldsymbol{y_0} - \boldsymbol{x_0}\right| \le V.$$

$S_y$ is a translated version of $S_x$. The fork centred on $\boldsymbol{x_0}$ is said to match that at $\boldsymbol{y_0}$ ($S_x$ matches $S_y$) if all the colour components of corresponding pixels are within threshold $\delta_k$,

$$\left|F_k(\boldsymbol{x}'_i) - F_k(\boldsymbol{y}'_i)\right| \le \delta_k, \quad k = r, g, b, \ i = 1, 2, ..., m. \tag{6}$$

All pixels ($N = V^2$) within the view radius are searched to find matches and the corresponding displacements are recorded as follows:

For the $j$th of $N_l < N$ matches define the corresponding displacement between $\boldsymbol{x_0}$ and $\boldsymbol{y_0}$ as $\boldsymbol{\sigma}_j^{t+1} = \left(\sigma_p, \sigma_q\right)$ where

$$\sigma_p = \left|x_{0p} - y_{0p}\right|, \quad \sigma_q = \left|x_{0q} - y_{0q}\right|, \tag{7}$$

and the cumulative displacements $\Delta$ and match counts $\Gamma$ as

$$\left.\begin{aligned} \Delta(\boldsymbol{x_0}) &= \Delta(\boldsymbol{x_0}) + \boldsymbol{\sigma}_j^{t+1} \\ \Gamma(\boldsymbol{x_0}) &= \Gamma(\boldsymbol{x_0}) + 1 \end{aligned}\right\} \; j = 1,\ldots,N_l < N, \tag{8}$$

where $N_l$ is the total number of matching forks and $N$ is the total number of matching attempts.

The displacement $\overline{\boldsymbol{\sigma}}_{\boldsymbol{x_0}}^{t+1}$ corresponding to pixel $\boldsymbol{x_0}$ averaged over the matching forks is

$$\overline{\boldsymbol{\sigma}}_{\boldsymbol{x_0}}^{t+1} = \frac{\Delta(\boldsymbol{x}_0)}{\Gamma(\boldsymbol{x}_0)}. \tag{9}$$

This process is carried out for every pixel $\boldsymbol{x_0}$ in the candidate motion region $R_t$. All internally mismatching forks $S_x$ with m = 2 at each pixel location are used for matching between the two frames. The displacements are saved in the motion vector map $O_{MV}$ and a copy in $O_{MV}^{'}$.

### 2.2. Motion segmentation based on region growing and matching

A region growing and matching process [15] is applied to obtain homogeneous regions with motion information. The motion vectors generated in the previous section tend to be associated with salient regions such as leading and trailing edges of moving objects; non-salient homogeneous regions are not assigned motion vectors and for this reason in the second stage a region growing algorithm is introduced which infers motion in these homogeneous regions. First homogeneous regions are identified. Then the position of the largest motion vector is taken as a seed for region growing and the value of this vector is assigned to pixels in the homogeneous region if this translation would lead to a pixel match in the next frame. This is repeated for the same homogeneous region to allow a different motion vector to be assigned to the remaining part of the same homogeneous region to obtain a match with the next frame. Regions which are changing shape would be affected by this process.

Seed motion vectors are rejected if their locations are not present in a difference frame between the current and next frame. This eliminates the spurious analysis of stationary objects not present in the reference frame.

### 2.3. Object tracking

Once the salient motion regions are obtained after growing, their corresponding centroids are then linked across multiple frames and used for tracking in the video sequence. Regions are linked if they overlap each other between successive frames. The areas of regions in each frame are normalized according to perspective in the image and ordered in descending order according to their sizes. A region is selected for tracking if its size is larger than a certain threshold K. K is set to 1000 for a 720x576 image and scaled according other image sizes. Tracking trajectories are plotted separately in the x and y directions against the frame number.

## 3. RESULTS AND DISCUSSION

The attention based region growing algorithm is illustrated on various data including road scenes from an MPEG-7 traffic sequence [16] and a London Train Station pedestrian sequence from PETS2006 [17]. The parameters of all experiments are $\varepsilon = 1$, $m$ = 2, $\boldsymbol{\delta}$ = (40,40,40), $T$ = 90. 100 frames from each video are used for tracking. The varying parameters are the view radius, $V$ and K. $V$ is selected according to the maximum velocity expected in the clip.

### 3.1. Traffic sequence

A traffic sequence of frame size 352x288 (Figure 1) was analysed with results shown in Figure 2. The reference frame was obtained by averaging over 1418 frames. Areas of candidate motion were obtained by taking the difference between each frame and the reference frame. The network of motion trajectories arises from the dividing and rejoining of homogeneous regions as the motion progresses across frames. $V$ was set to 20 this corresponding to the maximum expected velocity of the objects. Regions containing more than 250 pixels were tracked.

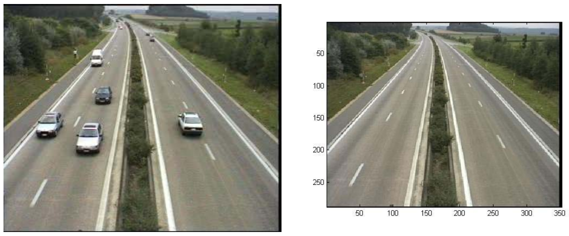
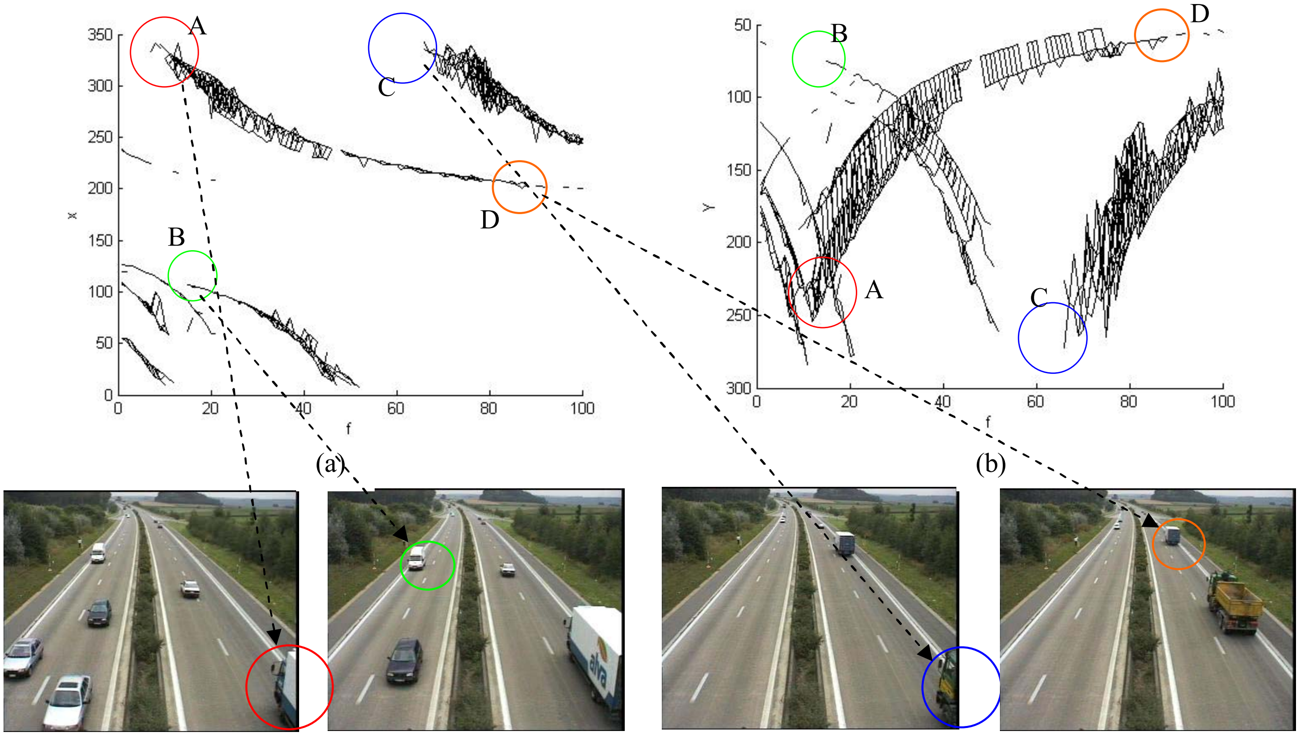


**Fig.1.** First frame (left) and reference frame (right)



(a)

(b)

(c)

(d)

(e)

(f)

**Fig.2.** (a) X-frame plot; (b) Y-frame plot; (c) frame 7; (d) frame 14; (e) frame 66; (f) frame 85

Figures 2(a) and 2(b) show the X and Y trajectories against frame number for 7 vehicles. The red circle (A) indicates the point when the white truck enters the scene at frame 7; the green circle (B) indicates the point where the white van starts being tracked at frame 14; the blue circle (C) indicates the point when the final truck enters the scene at frame 66; the orange circle (D) indicates the point where the white van ceases being tracked. The motion estimation process for each frame takes approximately 30 seconds while the growing process takes around 10 seconds running in C++ on a 2.8 GHz machine with 512 MB RAM.

### 3.2. London train station

The tracking results are generated from a London Train Station sequence of frame size 720x576. The reference frame was obtained by averaging over 1000 frames taken from the video. Motion trajectories in the x direction for each pedestrian are plotted in Figure 4. *V* was set to 15 this being the maximum expected object velocity in this scene. The average times for motion estimation and region growing are 20 seconds and 30 seconds respectively for each frame.

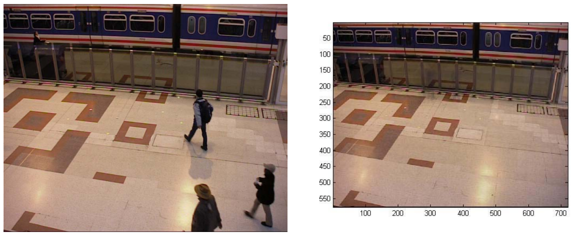


**Fig.3.** First frame (left) and reference frame (right)

Two pedestrians occlude each other at frames 69 (B) and 88 (C). The effect of shadows is indicated by a red circle at frame 50 (A) where the tracking is disturbed. Tracking stopped for the backpack pedestrian when he ceased moving around frame 40 (D). Three frames and the corresponding region growing maps are also shown in Figure 5 to illustrate these effects. Shadows from both pedestrians overlap each and lead to the linking of the two region centroids.

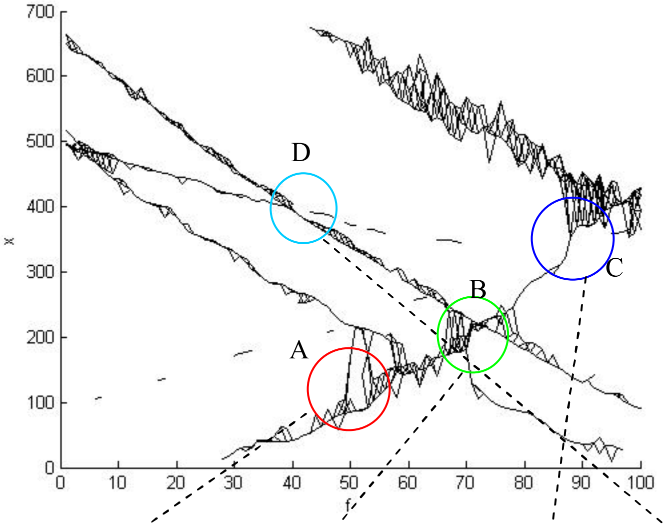
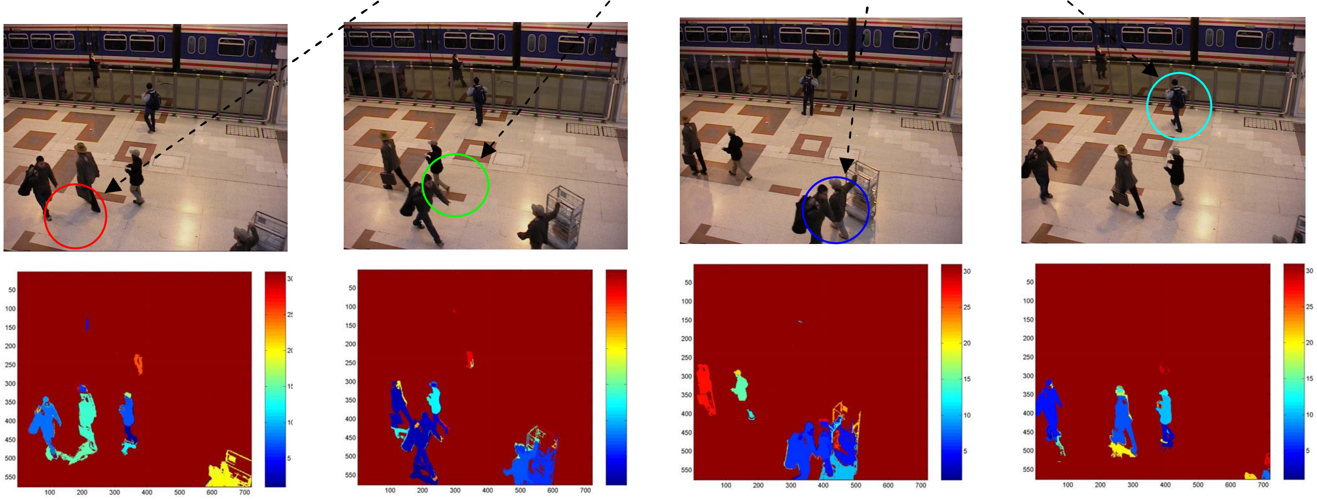


**Figure.4.** X-frame plot



(a) (b) (c) (d)

**Fig.5.** (a) frame 50 and its region map; (b) frame 69 and region map; (c) frame 88 and region map; (d) frame 40 and region map

## 4. CONCLUSIONS AND FUTURE WORK

A saliency based object tracking method has been proposed. The framework includes: an attention based approach that extracts object displacement between frames by comparing salient regions; a region growing technique that classifies motion regions according to motion information extracted and a simple matching process that assigns motion vectors to the classified regions; finally a tracking process which links region centroids between frames to form motion trajectories.

The tracking method was illustrated on various video data with a stationary background in both indoor and outdoor scenes. The method does not require a training stage or prior object models.

More accurate object tracking may be obtained by applying a shadow identification technique. However, future work is aimed at analyzing the network of motion trajectories to obtain more detailed object motion information which may also reveal distinguishing properties for shadows and objects. Tracking through occlusions will be developed. The proposed method will be compared with conventional tracking techniques such as mean-shift and particle filtering approaches. In addition more precise evaluation will be carried out using ground truth data.

## 5. ACKNOWLEDGEMENT

The project is sponsored by European Commission Framework Programme 6 Network of Excellence MUSCLE (Multimedia Understanding through Semantics, Computation and Learning) [18].

## 6. REFERENCES

[1] C. Stauffer and W.E.L Grimson, "Adaptive background mixture models for real-time tracking," in Proc. of CVPR, Ft. Collins, CO, USA, vol. 2, pp. 246-252, June 23-25, 1999.
[2] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in Proc. of CVPR, Hilton Head, SC, USA, vol. 2, pp.142-149, June 13-15, 2000.
[3] J. Shi and C. Tomasi, "Good features to track," in Proc. of CVPR, Seattle, WA, USA, pp. 593-600, June 21-23, 1994.
[4] M. Niethammer and A. Tannenbaum, "Dynamic geodesic snakes for visual tracking," in Proc. of CVPR, Washington, DC, USA, vol. 1, pp. 660-667, June 27-July 2, 2004.
[5] M. Isard and J. MacCormick, "BraMBLe: a Bayesian multiple-blob tracker," in Proc. of ICCV, Vancouver, Canada, vol. 2, pp. 34-41, July 7-14, 2001.
[6] T. Veit, F. Cao, and P. Bouthemy, "Probabilistic parameter-free motion detection," in Proc. of CVPR, Washington, DC, USA, vol. 1, pp. 715-721, June 27-July 2, 2004.
[7] M.J. Black and A.D. Jepson, "Estimating optical flow in segmented images using variable-order parametric models with local deformations," IEEE Trans. on PAMI, vol. 18, Issue 10, pp. 972-986, Oct. 1996.
[8] M.J. Black and P. Anandan, "The robust estimation of multiple motions: parametric and piecewise-smooth flow fields," CVIU, vol. 63, Issue 1, pp. 75-104, 1996.
[9] P. Viola, M.J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in Proc. of ICCV, Nice, France, vol. 2, pp. 734-741, Oct. 2003.
[10] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. on PAMI, vol. 20, Issue 11, pp. 1254-1259, Nov. 1998.
[11] L. Itti and P. Baldi, "A principled approach to detecting surprising events in video," in Proc. of CVPR, San Diego, CA, USA, vol. 1, pp. 631-637, June 2005.
[12] F.W.M. Stentiford, "An estimator for visual attention through competitive novelty with application to image compression," Picture Coding Symposium, Seoul, pp. 101-104, April 2001.
[13] L. Wixson, "Detecting salient motion by accumulating directionally-consistent flow," IEEE Trans. on PAMI, vol. 22, No. 8, pp. 774-780, Aug. 2000.
[14] S. Zhang and F.W.M. Stentiford, "Motion detection using a model of visual attention," in Proc. of ICIP, San Antonio, USA, pp. 513-516, Sept. 16-19, 2007.
[15] S. Zhang and F.W.M. Stentiford, "Motion segmentation using region growing and an attention based algorithm," European Conference on Visual Media Production, London, UK, Nov. 2007.
[16] MPEG-7 Content Set, Atlantic City, USA, Oct. 1998.
[17] Performance Evaluation of Tracking and Surveillance (PETS), http://ftp.pets.rdg.ac.uk.
[18] Multimedia Understanding through Semantics, Computation and Learning, 2005. EC 6th Framework Programme, FP6-507752, http://www.muscle-noe.org/.