

# Towards automated wide area visual surveillance: tracking objects between spatially-separated, uncalibrated views

R. Bowden and P. KaewTraKulPong

**Abstract:** This paper presents a solution to the problem of tracking intermittent targets that can overcome long-term occlusions, as well as movement between camera views. Unlike other approaches, our system does not require topological knowledge of the site or labelled training patterns during the learning period. The approach uses the statistical consistency of data obtained automatically over an extended period of time rather than explicit geometric calibration to automatically learn the salient reappearance periods for objects. This allows us to predict where objects may reappear, and within how long. We demonstrate how these salient reappearance periods can be used with a model of physical appearance to track objects between spatially separate regions in single and separated views.

## 1 Introduction

Intelligent visual surveillance is an important application area for computer vision. In situations where networks of hundreds of cameras are used to cover a wide area, the obvious limitation is the user's ability to manage such vast amounts of information. For this reason, automated tools that can generalise about activities or track objects are an important tool to the operator. Key to the user's requirements is the ability to track objects across (spatially-separated) camera scenes. However, extensive geometric knowledge about the site and camera position is typically required. Such an explicit mapping from camera to placement is infeasible for large installations as it requires that the operator know which camera to switch to when an object disappears. To further compound the problem the installation costs of CCTV systems outweigh those of the hardware. This means that geometric constraints or any form of calibration (such as that which might be used with epipolar constraints) is simply not realistic for real world installation. The algorithms cannot afford to dictate to the installer. This work attempts to address this problem and outlines a method to allow objects to be related and tracked across cameras without any explicit calibration, be it either geometric or colour.

Algorithms for tracking multiple objects through occlusion normally perform well for relatively simple scenes where occlusions by static objects and perspective effects are not severe. For example, Fig. 1 shows a typical

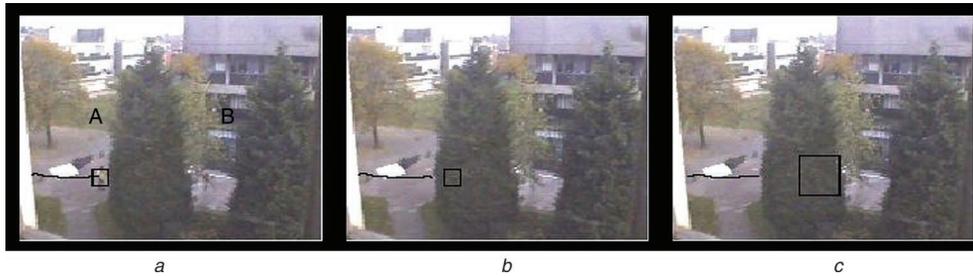
surveillance camera view with two distinct regions **A** and **B** formed by the presence of a static foreground object (tree) that obscures the ground from view.

In this paper, regions or subregions are defined as separated portions grouped spatially within an image. A region can contain one or more paths that may cover an arbitrary number of areas. Tracking an object across regions in the scene, where the geometric relationship of the regions cannot be assumed, poses similar challenges to tracking an object across spatially-separated camera scenes (again where the geometric relationship among the cameras are unknown). In a single view, the simplest solution to this problem is to increase the allowable number of consecutive frames that a target persists with no observation before tracking is terminated. This process is shown in Fig. 1 using a Kalman filter as a linear estimator.

By delaying the tracking termination, both the deterministic and the random components within the dynamics of the Kalman filter propagate over time [1], increasing the uncertainty of the predicted area in which the target may reappear (as shown in Fig. 1c). This increases the chance of matching targets undergoing long occlusions, but also increases the chance of false matches. In situations where linear prediction cannot be assumed (for example, the pathway changes direction behind large static objects), this will result in a model mismatch and the kinematic model assumed in most trackers will provide incorrect predictions. Furthermore, this type of approach cannot be extended to multiple cameras without an explicit calibration of those cameras.

## 2 Previous work

An approach often used to tackle object tracking across multiple cameras is to ensure that some overlap within the field of view of the cameras is available. An example, is the pilot military system [2, 3] where a bird's-eye camera view is used to provide a global map. The registration of a number of ground-based cameras to the global map allows tracking to be performed across spatially-separated camera scenes. The self-calibrated multiple camera system demonstrated in [4] assumes partially-overlapping cameras. Epipolar geometry, landmarks, and a target's visual



**Fig. 1** Tracking reappearance targets

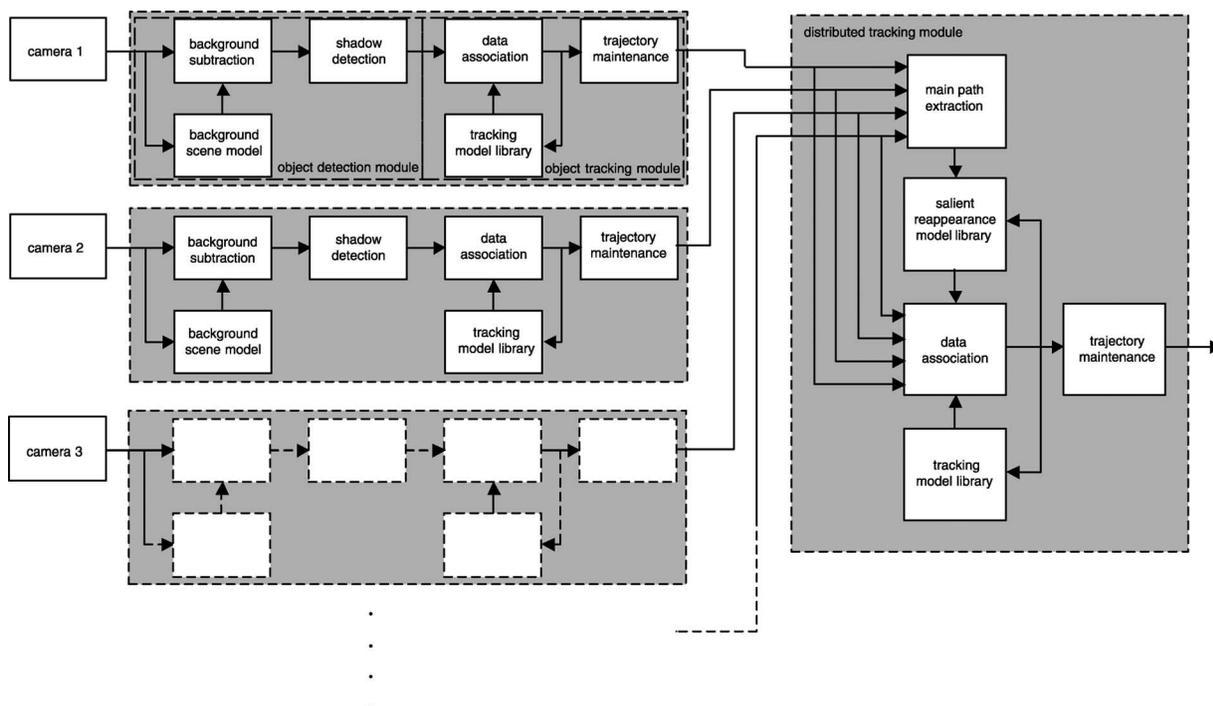
- a Target is being tracked
- b Target is occluded but the search continues
- c Uncertainty propagation increases over time

appearance are used to facilitate the tracking of multiple targets across cameras and to resolve occlusions.

Tracking across non-overlapping views has been of recent interest to many researchers. Huang and Russell [5] developed a system to track vehicles on a highway across spatially-separated cameras. In their system, knowledge about the entrance and exit of the vehicles into the camera views must be provided along with transition probabilities. Kettner and Zabih [6] presented a Bayesian framework to track objects across camera views. The user supplies topological knowledge of usual paths and transition probabilities. Javed *et al.* [7] presented a more general solution to the problem by providing an update of inter-camera parameters and appearance probabilities. However, their method assumes initial correspondences of those parameters. Our method makes use of data obtained automatically to discover such relationships between camera views without user intervention.

In accumulating evidence of patterns over time we expect to discover common activities. These patterns can be modelled in a number of ways. They can be used to classify sequences as well as individual instances of a sequence or to discover common activities [8]. Howarth and Buxton [9] introduced a spatial model in the form of a hierarchical structure of small areas for event detection in traffic surveillance. The model is constructed manually from tracking data. Fernyhough *et al.* [10] used tracked data to

build a spatial model to represent areas, paths and regions in space. The model is constructed using a frequency distribution collected from a convex-hull binary image of the objects. Thresholding of the frequency distribution, filters out low distribution areas, *i.e.* noise. Johnson and Hogg [11] used flow vectors, *i.e.* the 2-D position and velocity, collected from an image sequence over an extended period to build a probability distribution of the targets moving in the image. A neural network with competitive layers is used to quantise the data and represent the distribution. Its use is to detect atypical events which occurred in the scene. Makris and Ellis [12] used spline representations to model common routes and the activity of these routes. Entry and exit points and junctions are identified as well as their frequencies. Nair and Clark [13] used extended data to train two hidden markov models (HMMs) to recognise people entering and exiting a room in a corridor. Uncommon activity is identified as break-in by calculating the likelihood of the trajectories of the HMMs and comparing with a predefined threshold. Stauffer [14] used online vector quantisation, described in [11], to quantise all target information including position, velocity, size, and binary object silhouette into 400 prototypes. They performed inference on the data to obtain a probability distribution of the prototypes encoded in a co-occurrence matrix. A normalised cut [15] is then performed on the matrix which results in grouping



**Fig. 2** System overview

similar targets in terms of the above features in a hierarchical form.

### 3 Overview

Figure 2 gives a general overview of the system with the per camera elements separated from those that correlate objects between cameras and regions. Each camera is first fed to an object detection module. Here a background scene model is maintained and used to segment foreground objects on the fly. This basic segmentation is then refined by identifying areas of misclassification due to shadows using the chromaticity of individual pixels. Following this, objects are passed to the object tracking module where data association attempts to provide consistent temporal labelling of objects as they move. This is done by maintaining a library of currently-tracked objects that summarises all measurements into the relevant motion and appearance models. The output of this module are trajectories that exhibit temporal and spatial consistency. Specific details on these stages of tracking are presented in [16].

Following this, the resulting tracked objects in each camera are passed to the distributed tracking module, which attempts to connect seemingly unrelated trajectories (both spatially and temporally) into consistent object labels across regions and cameras. The first step is to extract main paths. This involves an unsupervised clustering of trajectories; grouping like trajectories, discarding spurious detections and any trajectories which have insufficient supporting evidence to provide reliable results. Following this, a model library of salient reappearance periods is constructed between all main paths and used in the data association stage to link trajectories that are spatially and temporally distinct from each other. The output of this module is the consistent labelling of objects despite occlusions or disappearances between cameras.

### 4 Relating possible reappearing targets

In order to identify the same target reappearing after a long occlusion, features and/or characteristics of object similarity must be identified. The properties of reappearing targets are assumed as follows:

- a target should disappear from one area and reappear in another within a certain length of time
- the reappearing target must occur only after that target has disappeared. (There is no co-existence of the same target at any instance. This is according to the spatially-separated assumption. For overlapped camera views, this assumption may not be applied.)
- frequently occurring disappearances or reappearances will form consistent trends within the data

An example of plausible matching can be seen in Fig. 3. In this Figure the black bar indicates the time line of the target of interest. Grey bars are the targets that can make possible matches to the target of interest whereas white bars represent targets whose matching with the target of interest are illegal owing to breaches of the above assumptions.

If targets are moving at similar speeds along a path occluded by large static objects (such as the tree in Fig. 1), over a period of time, there should be a number of targets that disappear from a specific area of region A and reappear in another area of region B within a temporal window. This can be called the salient reappearance period between the two areas. Both areas can be considered to be elements of the same path even though they are in different regions. The reappearance periods of these targets should be similar compared with random appearances of targets between other elements.

We start by automatically collecting data from our tracking algorithm in a single camera. The data consist of tracking details of targets passing into the field of view of the camera. The tracker is based upon modelling the background colour distribution on a per pixel basis using a Gaussian mixture model learnt using an online approximation to expectation maximisation using a similar framework to that originally proposed by Stauffer and Grimson [8]. This provides a binary segmentation of foreground and background delineation. Colour is then used to remove shadows by relaxing the model constraints upon intensity but not chromaticity. Foreground objects are extracted via connected component analysis and colour and shape used with a Kalman filter to track objects as they move. Occlusions are resolved using a stochastic sampling search

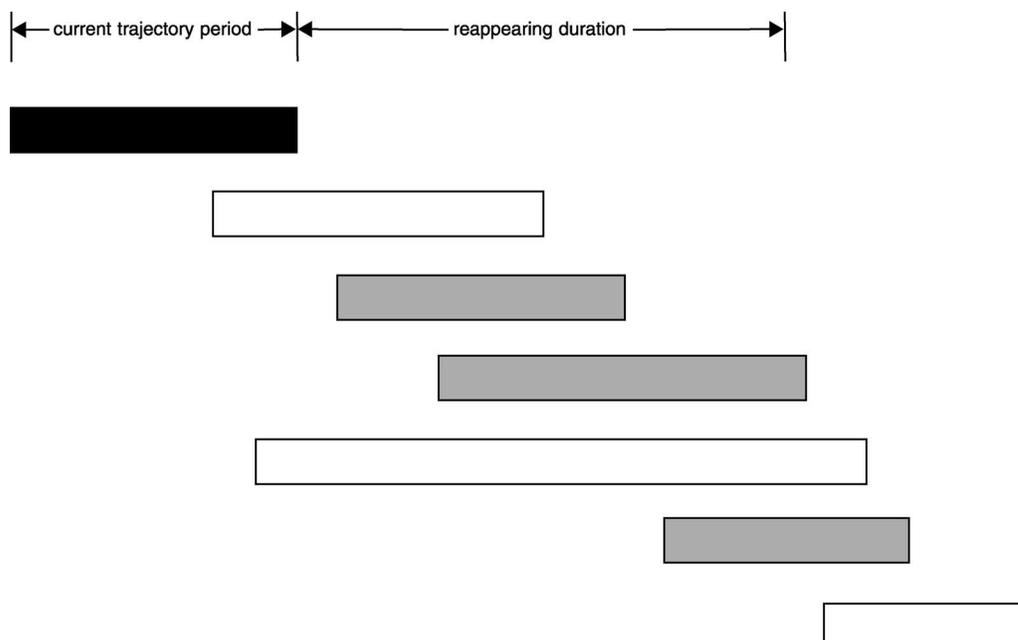


Fig. 3 An example timeline of plausible matching

to match partial objects to previous histories. Full details of the base tracking algorithm are given in [16].

Figure 4 shows trajectories of all objects collated. Notice that, owing to the occlusion of the tree, two separate regions are formed where no correspondence is known about the



Fig. 4 Trajectory data of the training set

relationship between them. The linear dynamics of the tracker are insufficient to cope with the occlusion.

The goal is to learn some reappearance relationship in an unsupervised fashion that allows objects that disappear to be located and tracked successfully when they reappear. This we call the salient reappearance period, and its construction is divided into two steps:

1. Extracting dominant paths: trajectories in each subregion are classified into a number of paths. Only 'main' paths that consist of a large number of supporting trajectories are extracted.
2. Extracting salient reappearance periods among dominant paths: In this stage, a set of features common to reappearing paths is introduced. The features allow possible reappearance targets to be calculated. A set of matches that show outstanding reappearance periods are chosen to train the matching models in the training phase.

#### 4.1 Colour similarity

We chose to use a colour histogram approach to describe the colour fingerprint of an object as it is simple and efficient to compute. Through quantisation it also allows some invariance to changes in colour appearance. A similarity

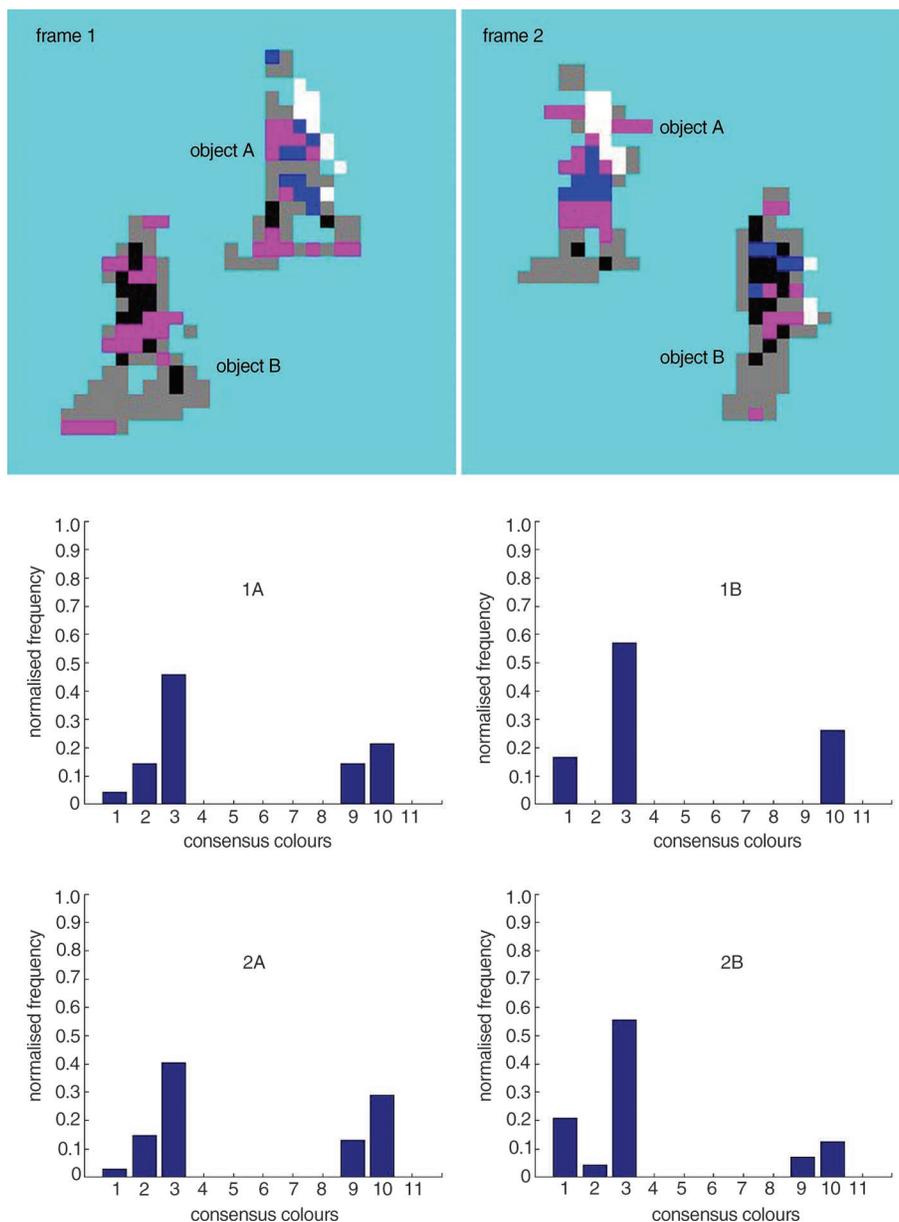


Fig. 5 Example colour descriptors for 2 objects over time

metric based upon histogram intersection provides a crude estimate of object similarity. Three colour spaces were investigated; RGB, HSL, and consensus–colour conversion of Munsell colour space (CCCM) as proposed by Sturges *et al.* [17]. A series of simple tests looked at the colour consistency of these methods for objects moving within a single image and also across images. We also investigated various levels of quantisation for both RGB and HSL. Our tests showed RGB to provide consistent results within a single image but poor performance (not unsurprisingly) across cameras that are not colour calibrated. HSL using 8-8-4 quantisation (as suggested in [18]) performed well in both tests but CCCM provides superior results intra–frame while retaining comparable results to HSL inter–frame. In-depth details of these tests are omitted owing to space limitations. CCCM breaks down all colours into 11 basic colours, the exact membership was determined experimentally in a physiological study where human categorisation was used to learn the membership. This coarse quantisation provides consistent intra camera labelling without colour calibration relying on the perceptual consistency of colour, *i.e.* if a red object is perceived as red in both images CCCM will provide a consistent label.

Figure 5 shows two images from a sequence, labelled frame 1 and frame 2. These frames contain the same two objects (A and B) at different times. Although spatially the appearances of the objects differ, the CCCM colour histograms are relatively consistent for each.

#### 4.2 Path extraction

Paths can be represented effectively as a group of trajectories between two areas. However, some trajectories may begin and end in the same area. Paths of this type must be divided into different groups. This is done using a normalised cut [15] of the 4-D motion vectors formed through the concatenation of position and velocity ( $x, y, dx, dy$ ). All paths that contain more than two percent of the total trajectories are kept for further processing and are shown in

Fig. 6. Paths with a small number of supporting trajectories are sensitive to noise and can produce erroneous trends. They are therefore discarded. The Figure shows how the normalised cut has broken trajectories down into groups of like trajectories.

#### 4.3 Linking paths

Paths are linked by building a fuzzy histogram of all possible links among trajectories of a pair of main paths from different regions within a period named the ‘allowable reappearance period’. The bin  $\tau_t$  of the histogram is calculated from

$$\tau_t = \sum_{\forall i,j} H_{ij}; (t_i^{end} - t_j^{start}) < t \quad (1)$$

where  $t_j^{start}$  and  $t_i^{end}$  are the time instances that target  $i$  starts and ends respectively.  $H_{ij}$  is the result from histogram intersection between the colour histogram of target  $i$  and that of target  $j$ .

$$H_{ij} = \sum_{k=1}^{11} \min(B_{ik}, B_{jk}) \quad (2)$$

The colour appearance histograms,  $B_j = \{B_{j1}, B_{j2}, \dots, B_{j11}\}$ , are constructed by taking the constituent pixels of the object and categorising each pixel as one of the 11 basic colours using CCCM (as discussed previously), providing consistent labelling of colour across cameras without the burden of colour calibration.  $H_{ij}$  is within the range of 0 to 1 with 1 corresponding to a perfect match.

An example of the fuzzy histogram of reappearance periods within 60 seconds from sample path A to B (Fig. 7) is shown in Fig. 8. The histogram bin size was set at one second and the frequency of the bin was the summation of colour similarity scores using CCCM. A full discussion of the allowable reappearance period, fuzzy frequency and the choice of parameters used is given in [19].

Using the linking scheme described previously on every pair of main paths between two regions (with only main

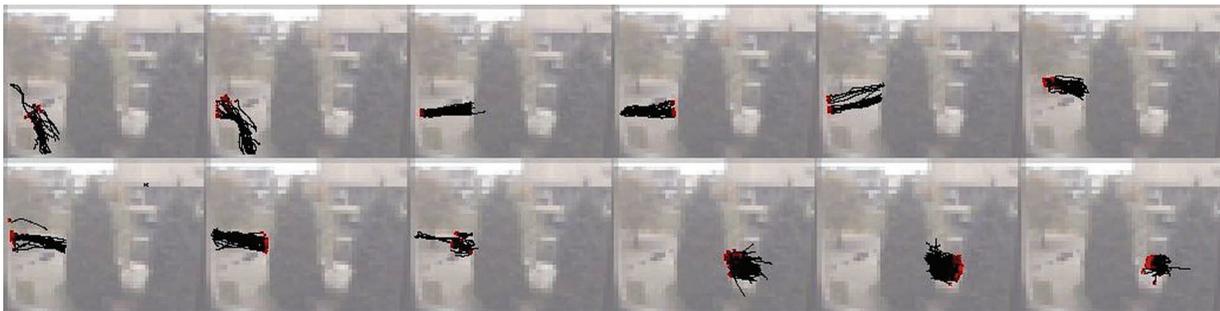
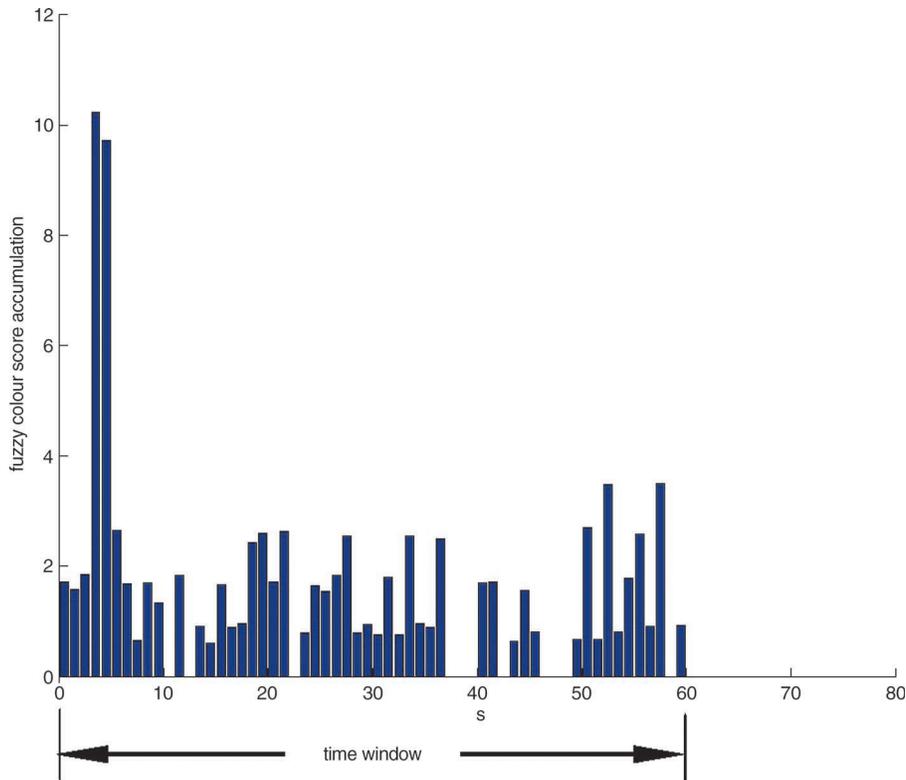


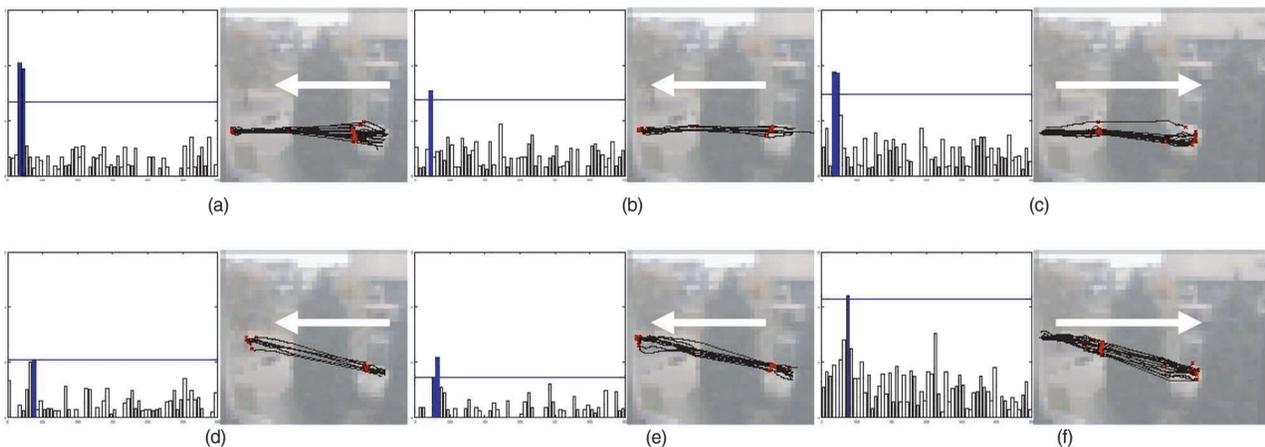
Fig. 6 Main paths in the first region



Fig. 7 An example of a pair of possible paths to be matched



**Fig. 8** Fuzzy histogram of reappearance periods from sample path A to B with an allowable reappearance period of 60 s



**Fig. 9** Extracted salient reappearance periods among main paths between different regions

paths from different regions permitted for the matches), a number of possible matches produces salient reappearance periods. Figure 9 shows the results. Two histograms for every pair of paths are produced. The one with the maximum peak is selected. To check the validity of the outstanding peak it must exceed four times the noise floor level. The noise floor level was found by taking the median from non-empty bins of the histogram. Unimodality is assumed. Therefore, a single peak is detected based on the maximum area under the bins that pass the noise level. This could, of course, be used in a particle filter framework should unimodality not be sufficient. However, our work thus far has not found it necessary. The histograms are presented with the corresponding linked trajectories in Fig. 9 with white arrows showing the direction of motion. The detected bins are shown with solid bars on each histogram, the noise level is also plotted in each histogram.

With the data collected automatically from the last process, a reappearing-target model can be formed for each pair of detected main paths. A simple model is calculated

from the training data as follows. In each pair of detected main paths, the reappearance periods  $\{r_1, r_2, \dots, r_N\}$  between a pair of paths is represented compactly by their mean  $\mu_r$  and standard deviation  $\sigma_r$ .

## 5 Path-based target recognition

Based on the model obtained in the training phase, target recognition can be performed. Target recognition is based on the same linking scheme. The process can be performed online using the same principle as multiple hypothesis tracking which, select the best hypothesis within the reappearance period, or offline after a batch collection of trajectories within the allowable reappearance period.

Similarity based on the salient reappearance periods is calculated in the following manner. First, the standard (zero mean and unity variance) normal random variable of each reappearance period is calculated using  $z = \frac{r - \mu_r}{\sigma_r}$ . Then the standard normal cumulative probability in the

right-hand tail  $\Pr(Z > z)$  is determined using a look-up table. The similarity score is two times this value which gives the score in the range of 0 to 1. Another way of calculating this value is

$$2\Pr(Z > z) = 1 - \int_0^z f(x|1)dx \quad (3)$$

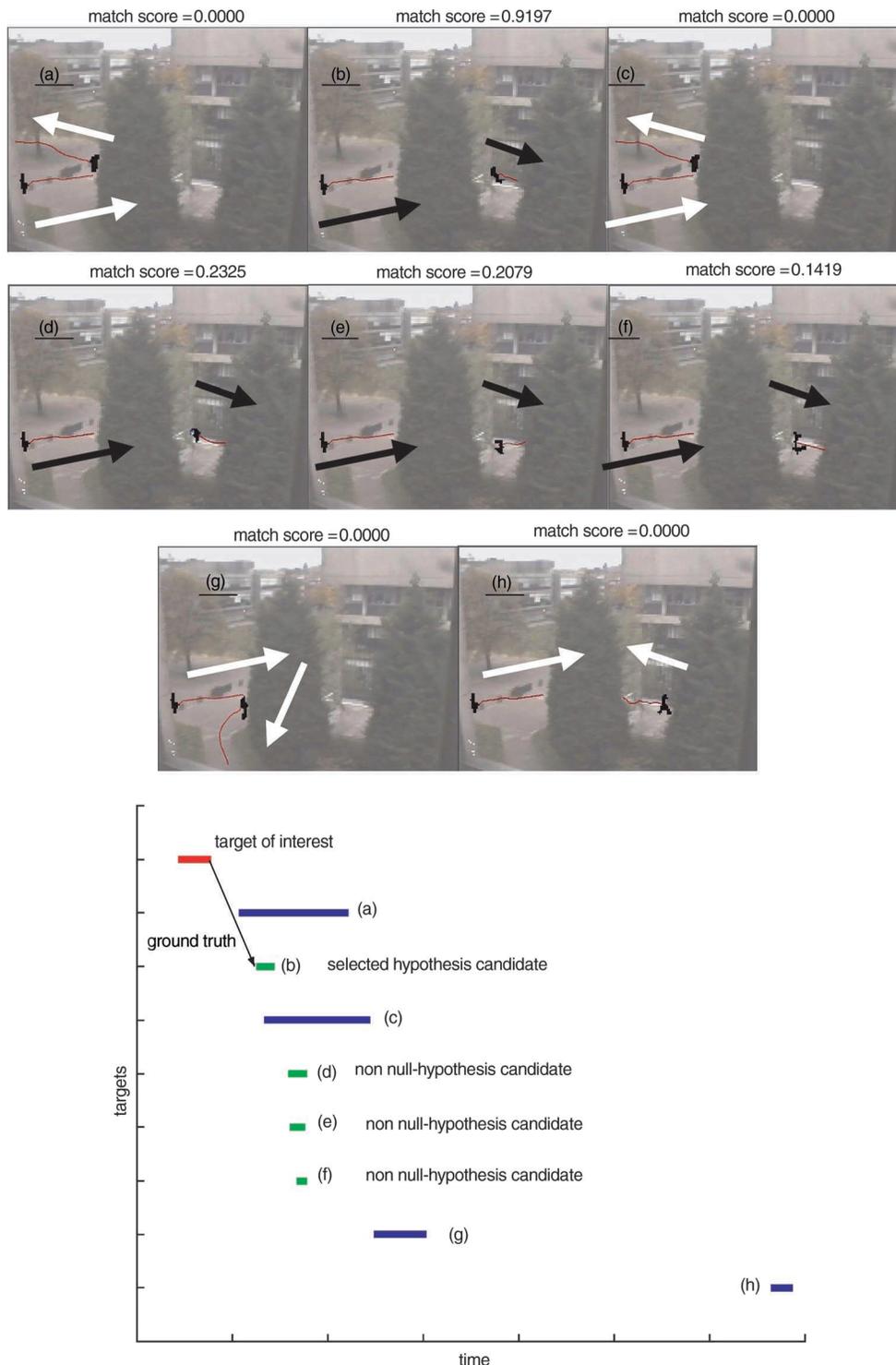
where  $f(x/v)$  is the chi-squared distribution with  $v$  degrees of freedom.

$$f(x|v) = \frac{x^{\frac{(v-2)}{2}} e^{-\frac{x}{2}}}{2^{\frac{v}{2}} \Gamma(\frac{v}{2})} \quad (4)$$

where  $\Gamma(v)$  is the Gamma function. However, a no match hypothesis or null hypothesis is also introduced, as it is

possible to have no link. Hypothesis testing for this null hypothesis is required before any classification is performed. A score of 0.001 was set for the null hypothesis which corresponds to the standard value  $z$  of 3.09. Any candidates which are not 'null hypotheses' are selected based on their maximum score. Online recognition selects the best candidate at each time instance within the allowable reappearance period. This allows the tracker to change its link each time a better hypothesis is introduced. Batch recognition, alternatively, collects all trajectories within the allowable reappearance period and performs the classification based on the whole set.

In the first experiment, the training data was collected automatically by our target tracking algorithm [16] over an extended period of time constituting 1009 individual



**Fig. 10** A example of recognising reappearing target between different regions and the associated time line

trajectories. An unseen test set of 94 trajectories was collected and hand labelled as ground truth. The recognition process is performed off-line by collecting the whole set of admissible targets according to the rules described in Section 3. An example of the recognition with similarity score and the corresponding time line chart is shown in Fig. 10. The motion of the objects is highlighted on the Figs. where white arrows denote implausible motions and black

arrows non null (plausible) hypotheses. The red line in the time line is the target of interest, while the green lines are the non null hypothesis candidates with plausible matches. The arrows depict the ground truth. It can be seen that target (b) was classified correctly to the ground truth as it obtained the best score during the recognition process.

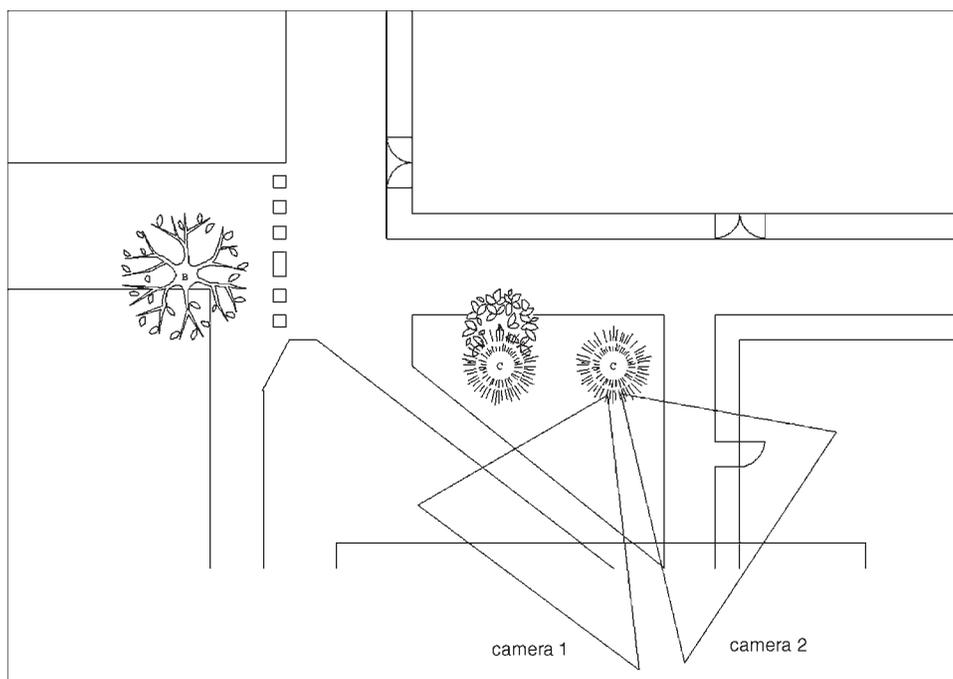
A summary of the recognition on the whole test set is provided in Table 1. Since the model is based on the trend in a pair of main paths, if the target uses uncommon paths that have no trend in the data set, the target cannot be recognised. This accounts for 2 out of the 6 misses. One of these misses was caused by a person who changed his mind during his disappearance and walked back to the same path in the same area, while the other was a result of camouflage at the beginning of the trajectory.

**Table 1: Matching results of an unseen data set in a single camera view**

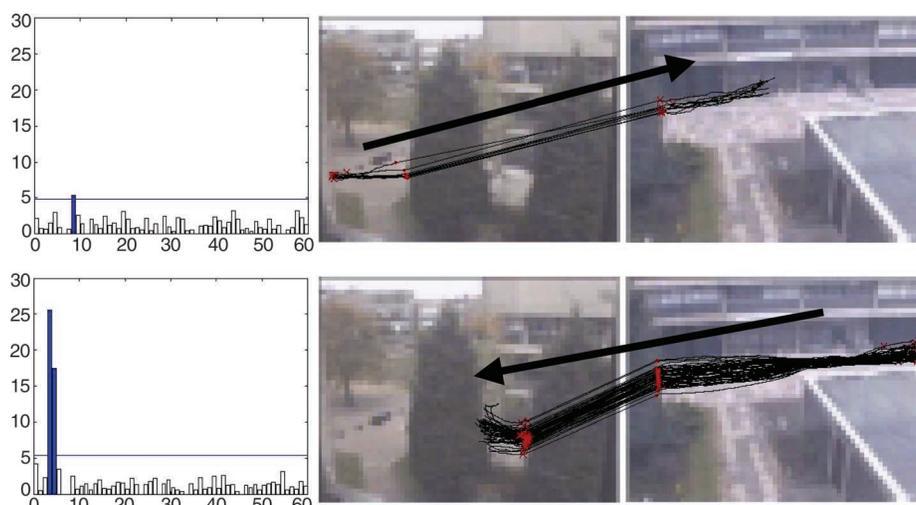
Items	Trajectories	%
Total training set	1009	
Total test set	94	100
Correct matches	84	89.36
Total incorrect matches	10	10.64
False detections	4	4.26
Misses	6	6.38
Misses (uncommon paths)	2	2.13

## 6 Learning and recognising trajectory patterns across camera views

The target recognition presented can be extended also to multiple cameras. The same process of trajectory data collection from two scenes with time synchronisation was



**Fig. 11** Site map and camera layouts for recognising reappearing targets across camera views



**Fig. 12** Examples of extracting salient reappearance periods between main paths in different regions across camera views



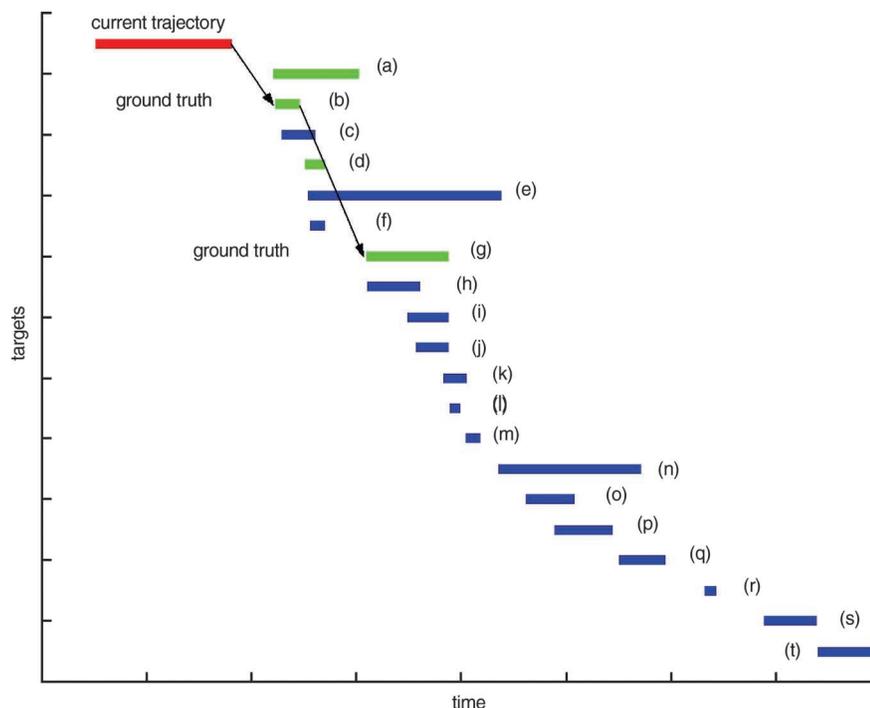
**Fig. 13** All possible links between paths in separate views

performed. The set-up of cameras is shown in Fig. 11. 2020 individual trajectories consisting of 56391 data points were collected for two cameras. An unseen test set of 133 trajectories was then hand labelled to provide ground truth.

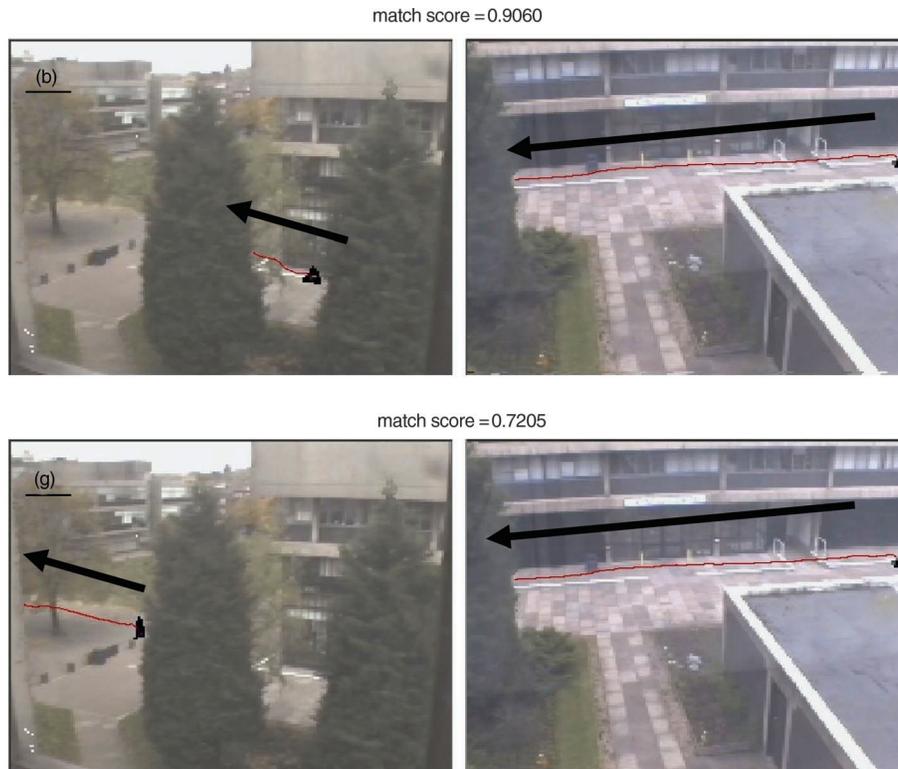
Figure 12 shows some examples of the salient appearing periods and trajectories extracted between camera pairs for main paths. Note that both of the examples are valid reappearing trends. However, the trend in the upper part of Fig. 12 is not as high owing to a small sample set in addition to an increased distance between the two main

paths. It should be noted if any two regions are physically too distant, the prominence of the salient reappearance period is reduced.

The results from pre-processing are then subjected to the same process as before and classification performed in the same way as for the single view. The test data were collected and hand labelled. It consists of 133 targets. For an example trajectory Fig. 13 shows all possible links to other paths within the time window. The corresponding time line chart is shown in Figure 14. Again, arrows assist in



**Fig. 14** Time line of an example of recognising reappearing target between different regions in spatially-separated cameras



**Fig. 15** Two most likely matches between trajectories

**Table 2: Matching results of an unseen set of 10 minutes across camera views**

Items	Trajectories	%
Total training set	2020	
Total test set	133	100
Correct matches	116	87.22
Total incorrect matches	17	12.78
False detections	8	6.02
Misses	9	6.77
Misses (uncommon paths)	2	1.50

visualising the direction of motion and colour coding is used to depict plausible (black) matches against null hypothesis (white) matches. The two highest candidates are reiterated in Fig. 15. In this example, candidate (b) is the best match and is selected. The second best, candidate (g), is also valid; however, it has a lower score owing to the flatter peak in its training set. This higher variance is caused by the greater distance between the two main paths that increases the effect of variance of the target speed. It is interesting to note that both of the highest matches are indeed correct as the current object becomes (b) and then after disappearing for a second time becomes (g). The approach naturally tries to predict as far into the future as possible. However, in practice, once the second disappearance has occurred a much stronger direct match between (b) and (g) would be used to perform the match. Matching results from the two scenes are shown in Table 2. Again a high number of correct matches are achieved, only slightly lower than the single camera results. Most notable is that using colour alone to match targets results in only around a 60% success rate but using colour to look for statistical trends spatially can achieve over 87%. More important is the number of false detections rather than misses, in terms of human evaluation it is this that is used to assess performance.

It can be seen from Tables 1 and 2 that the recognition rate of the proposed technique is high. However, the approach is still dependent on the assumption that the separated regions in the same or different views are close to each other. Future work will investigate the correlation between accuracy and distance for this technique. Although we demonstrate the approach here using two cameras, it is obvious to see how the approach could be extended to larger multiple camera installations and our future work will test this scalability.

## 7 Summary and conclusions

In this paper, an approach to recognising targets after occlusion is proposed. It is based on salient reappearance periods discovered from long-term data. By detecting and relating main paths from different regions and using a robust estimate of noise, salient reappearance periods can be detected with high signal-to-noise ratios. Offline recognition is performed to demonstrate the use of this extracted salient reappearance period and the appearance model to associate and track targets between spatially-separated regions. The demonstration is extended to regions between spatially-separated views with minimal modifications. As the underlying process of reappearance is not the salient reappearance time but the average distance between paths, the performance of this recognition process is degraded if the average distance between paths is increased. These issues need further investigation.

## 8 References

- 1 Blackman, S., and Popoli, R.: 'Design and analysis of modern tracking systems' (Artech House, Boston, MA, 1999)
- 2 Kanade, T., Collins, R., Lipton, A., Anandan, P., Burt, P., and Wixson, L.: 'Cooperative multi-sensor video surveillance'. DARPA Image Understanding Workshop, 1997, pp. 3–10
- 3 Kanade, T., Collins, R., Lipton, A., Burt, P., and Wixson, L.: 'Advances in cooperative multi-sensor video surveillance'. Darpa Image Understanding Workshop, 1998, pp. 3–24

- 4 Chang, T.H., Gong, S., and Ong, E.J.: 'Tracking multiple people under occlusion using multiple cameras'. BMVC00, 2000
- 5 Huang, T., and Russell, S.: 'Object identification in a bayesian context'. IJCAI97, 1997
- 6 Kettner, V., and Zabih, R.: 'Bayesian multi-camera surveillance'. CVPR99, 1999
- 7 Javed, O., Rasheed, Z., Shafique, K., and Shah, M.: 'Tracking across multiple cameras with disjoint views'. ICCV03, 2003
- 8 Stauffer, C., and Grimson, W.E.L.: 'Learning patterns of activity using real-time tracking', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2000, **22**, (8), pp. 747–757
- 9 Howarth, R.J., and Buxton, H.: 'Analogical representation of space and time', *Image Vis. Comput.*, 1992, **10**, pp. 467–478
- 10 Fernyhough, J.H., Cohn, A.G., and Hogg, D.C.: 'Generation of semantic regions from image sequences'. ECCV96, 1996, pp. II:475–484
- 11 Johnson, N., and Hogg, D.: 'Learning the distribution of object trajectories for event recognition', *Image Vis. Comput.*, 1996, **14**, (8), pp. 609–615
- 12 Makris, D., and Ellis, T.: 'Finding paths in video sequences'. BMVC01, 2001
- 13 Nair, V., and Clark, J.J.: 'Automated visual surveillance using hidden markov models'. VI02, 2002, p. 88
- 14 Stauffer, C.: 'Automatic hierarchical classification using time-based co-occurrences'. CVPR99, 1999, pp. II:333–339
- 15 Shi, J., and Malik, J.: 'Normalized cuts and image segmentation', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2000, **22**, (8), pp. 888–905
- 16 KaewTraKulPong, P., and Bowden, R.: 'A real-time adaptive visual surveillance system for tracking low resolution colour targets in dynamically changing scenes', *Image Vis. Comput.*, 2003, **21**, (10), pp. 913–929
- 17 Sturges, J., and Whitfield, T.W.A.: 'Locating basic colours in the munsell space', *Color Res. Appl.*, 1995, **20**, (6), pp. 364–376
- 18 Black, J., Ellis, T.J., and Makris, D.: 'Wide area surveillance with a multi camera network'. IDSS-04 Intelligent Distributed Surveillance Systems, 2003, pp. 21–25
- 19 KaewTraKulPong, P.: 'Adaptive probabilistic models for learning semantic patterns'. PhD thesis, Brunel University, 2002