Intelligent Visual Surveillance: Towards Cognitive Vision Systems

Dimitrios Makris[*,1], Tim Ellis[1] and James Black[1,2]

[1]*Faculty of Computing, Information Systems and Mathematics, Kingston University, UK*

[2]*Ipsotek Ltd., London, UK*

Abstract: Automated visual surveillance systems are required to emulate the cognitive abilities of surveillance personnel, who are able to detect, recognise and assess the severity of suspicious, unusual and threatening behaviours. We describe the architecture of our surveillance system, emphasising some of its high-level cognitive capabilities. In particular, we present a methodology for automatically learning semantic labels of scene features and automatic detection of atypical events. We also describe a framework that supports learning of a wider range of semantics, using a motion attention mechanism and exploiting long-term consistencies in video data.

1. INTRODUCTION

Visual surveillance systems are widely used in public places. Traditional surveillance systems consist of cameras, storage devices, video monitors and security personnel. Security staff monitor the activity in the scene, watching for suspicious or threatening activities. In addition to online monitoring, post-examination of recorded video data may be required to identify suspicious persons, vehicles or events. Both tasks are tedious, as security staff need to identify specific and unusual events from a large number of very common and repetitive events. Unfortunately, human operators usually struggle to deal with the required huge cognitive overload, even for a small surveillance system of few cameras.

Current commercial surveillance systems make use of digital technology to capture, store and process video data. For example, Video Motion Detectors (VMDs) are able to automatically detect scene motion and send a notification signal to an operator. However, their operation is still primitive and not sufficiently discriminatory (e.g. in busy environments, motion is continuously detected).

In general, visual surveillance systems are required to minimise the role of human operators. More specifically, some of the requirements are automatic detection of suspicious events that eases online monitoring, and context-based databases of the observed events that facilitate conceptual querying and searching. Cognitive vision systems that evolve and adapt to their environments could potentially fulfil such requirements. These systems are expected to outperform the human operators, as they will be able to operate reliably and continuously, without the fatigue constraint.

Whilst significant research has been undertaken into the problem of low-level vision tasks (motion detection [1, 2], tracking [3-5], etc.), understanding and interpreting complex activities may require a deeper understanding of the events occurring within the video data, in order to generate the relevant information to an operator, filtering out the mundane activity.

For this reason, the research community's interest has shifted to high-level tasks like event analysis, activity analysis and behaviour analysis [6-9]. High-level analysis of the surveillance video context can use mathematical models that do not necessarily correspond to a human interpretation, e.g. Hidden Markov Models (HMMs) [10]. However, because the surveillance system is required to interact with its operators, a cognitive knowledge base is required that will be common to both the surveillance system and the human personnel. A common cognitive knowledge base would allow the surveillance system not only to "understand" the video context, but also to provide automatic textual description of the video context, or answer contextual queries from the operators [11]. It is obvious that such an approach would significantly extend the functionality and usability of surveillance systems.

Research effort has been invested in providing surveillance systems with a semantic context for the knowledge base and the mechanisms that will allow the system to "understand" the video content and interpreted in terms of the semantics [12]. Usually, the semantics and the interpretation rules are manually encoded into the surveillance systems, endowing them with the ability of "knowing" and "understanding", two of the three main features of cognitive vision [13]. However, considerably less research has been undertaken to provide surveillance systems with the third characteristic ability of cognitive vision systems: "learning" [14, 15].

Adding a learning ability to surveillance systems is not just a challenge towards a cognitive vision system, but also a practical requirement, because it enables the system to build its own knowledge base, automatically adjusted to its environment. Also, learning allows the knowledge base to adapt

*Address correspondence to this author at the Faculty of Computing, Information Systems and Mathematics, Kingston University, UK;
E-mail: d.makris@kingston.ac.uk

1874-110X/08

to changes of the environment. The practical consequence is manifested if we consider the large number of operational surveillance cameras (e.g. over 4 million surveillance cameras are installed in the UK) and the human effort required to manually enrich them with a knowledge base consistent with their environment.

This paper presents our intelligent surveillance system and focuses on its cognitive aspects. It discusses how semantic labels of static scene features can be automatically learnt. Our approach is to exploit a large number of observations derived by lower level modules (motion detection, motion tracking) over extended periods of time from an online surveillance system. Details of the motion detection, tracking and learning route models, entry/exit zones and stop zones can be found elsewhere [16-18]. Learning is performed by identifying spatially-related long-term consistencies in the data, using a motion attention mechanism. It also discusses how typical patterns of activity can be established and used for detecting suspicious events. The above methodologies are discussed in the context of both single and multiple camera systems.

Section 2 introduces the architecture of our intelligent surveillance system. Section 3 discusses the types of semantics that are required to describe the activities observed by a surveillance system. Section 4 presents how scene and activity models are learnt from observations. The activity/scene models are used to detect suspicious activities in section 5. Finally, section 6 summarises the conclusions and discusses possible extensions of our work.

2. BACKGROUND

This section presents the Kingston University Experimental Surveillance (KUES) system that we developed. The architecture of the system, based on a distributed multi-camera network of cooperating independent processors, is illustrated in Fig. (**1**). A motion attention mechanism is implemented by the motion detection and tracking modules, while high-level cognitive tasks are performed by the learning and understanding modules. Object classification, the interpretation of articulated human motion and the labelling of human actions [19, 20] are beyond of the scope of the system.

Each surveillance camera generates a video stream. The motion detection module [16] establishes a pixel-wise background model for each camera view and identifies pixels in each frame where motion is present (foreground), assuming that variations on the values of the pixels over time are caused by the motion of the targets. Then, the foreground pixels are segmented into Binary Large OBjects (Blobs).

Motion tracking [16] aims to provide one trajectory for each individual target that represents the time sequence of the target centroid positions, within the camera view. For this reason, motion tracking attempts to correspond blobs in consecutive frames and resolve ambiguities caused by misdetections or occlusions. Blobs are described by a set of characteristics (position, velocity, size, colour) that are used as matching criteria.

The 3D motion-tracking module aims to encode the complete history of individual targets moving within the

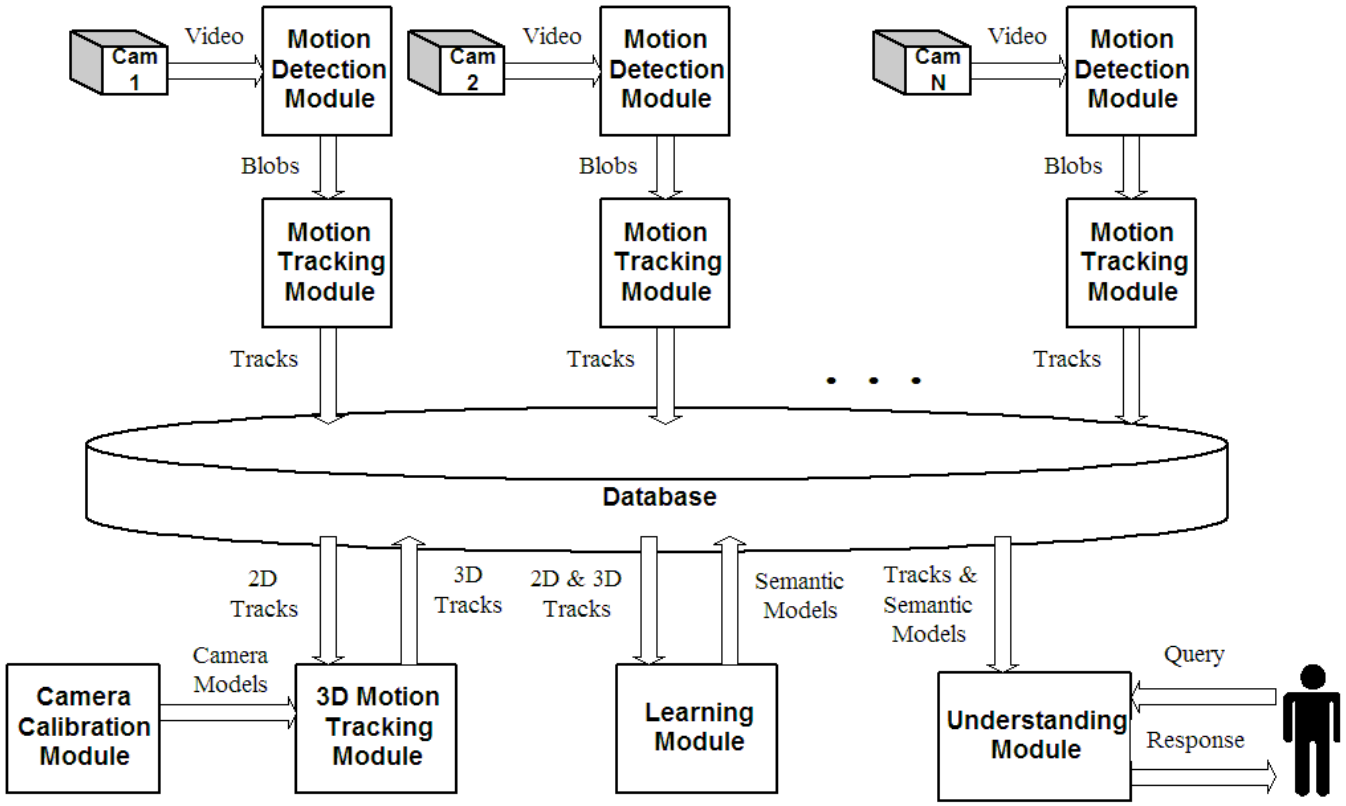

Fig. (1). Architecture of the KUES system.

entire region viewed by the camera network. Image based tracks provided by different cameras are combined according to geometric camera models obtained from a one-off calibration process. The 3D trajectories are usually expressed in terms of a common ground plane coordinate system. The benefit of using 3D trajectories is that they represent the motion history of targets in terms of the real scene (ground plane).

The learning module uses the 2D and 3D tracks to generate semantic scene and activity models. The scene models enable the system to "understand" the motion and respond to contextual queries made by the operator. Activity models are used to detect "suspicious" events that are not consistent with them.

The KUES system was operated and recorded data continuously over a period of several months. A hierarchical database [21] contains video recorded data (low level), trajectory representations (mid level) and conceptual descriptors (high level).

3. SEMANTIC DICTIONARY FOR VISUAL SURVEILLANCE OF VIDEO ANNOTATION

We propose a general scheme to describe activity observed by surveillance systems. Specifically, descriptions of observed activities can be based on a semantic dictionary that contain three main categories: a) moving "targets" (pedestrians or vehicles), b) actions [21, 23] and c) static features of the scene.

According to this scheme, moving targets perform actions within a scene described by static features, and interact either with other targets or with the static features of the scene. More formally, video annotation sentences are formed using the moving "targets" as subjects or objects of a sentence, the actions as verbs and the static features of the scene as objects or part of a locative adjunct or locative complementary.

For instance, textual descriptions like "Mr X enters the room from the door" and "A red car stops before the pedestrian crossing", contain all three types of semantics: moving targets like "Mr X" and "red car", actions like "enters", "stops" and static features like "door" and "pedestrian crossing". In the case that targets interact with other targets and/or static features (e.g: "Mr X handshakes Mr Y", and "Mr X uses the ATM machine") the proposed semantic dictionary seems adequate.

4. LEARNING AN ACTIVITY-BASED SEMANTIC SCENE MODEL

We aim to provide surveillance systems with the ability to automatically build a knowledge base of their environment. We employ unsupervised learning to exploit the vast amount of track data, obtained during extended periods of time.

The scene structure directly restricts or indirectly influences the way that targets act. Therefore, specific types of events may be associated with specific regions. For instance, roads constrain vehicles to move along specific lanes in a particular direction; gates and doors are related to entrance/exit events where targets will appear or disappear; bus stops indicate where people should wait for the bus to stop. Therefore, the proposed learning of spatially-related semantic labels that exploits the motion attention mechanisms of the KUES is actually a reverse engineering approach for identifying these regions (Fig. **2**).

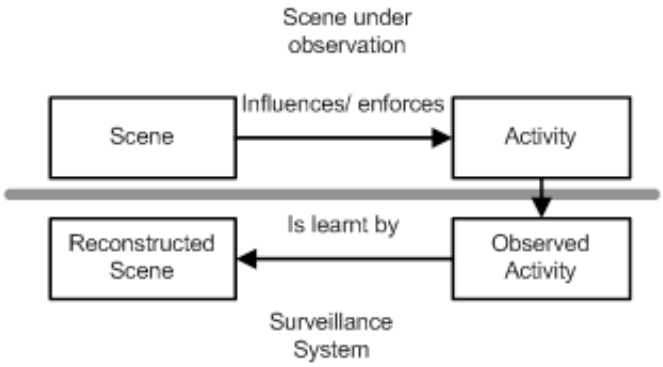

Fig. (2). Scene reconstruction using activity.

As a consequence, the semantics that can be learnt are activity-based. In [17, 18], we have proposed a scene model that describe semantic features such as routes, paths, junctions, entry/exit zones and stop zones. Fig. (**3**) represents a topographical view of the proposed model, while Fig. (**4**) depicts a topological view. Fig. (**5**) depicts a manually constructed model, overlaid on the image plane. Semantic modelling fulfils two requirements: a geometric description of the spatial extent of the scene features and quantitative representation of the related activity. While the first requirement is obvious for spatial-related features, the second requirement is necessary to derive models that can support a probabilistic interpretation of "understanding" activity.

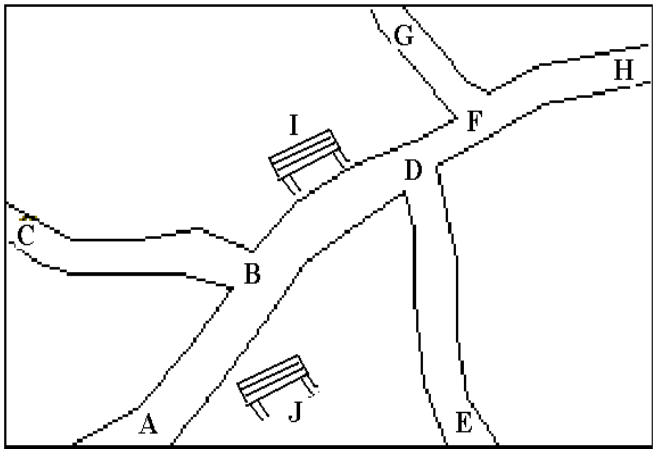

Fig. (3). Topographical representation of environment. Entry/exit zones (A, C, E, G, H), junctions (B, D, F), stop zones (I, J), paths (AB, CB, BD, DF, etc) and routes (ABDFH, ABC, EDFG, etc) are depicted.

Entry/exit zones are associated with instantaneous events of an object entering or exiting the scene and each event can be detected and localised. For each trajectory in the database, one entry and one exit event can be obtained. The set of entry/exit points can be modelled by a Gaussian Mixture Model (GMM) and learnt by an Expectation-Maximisation

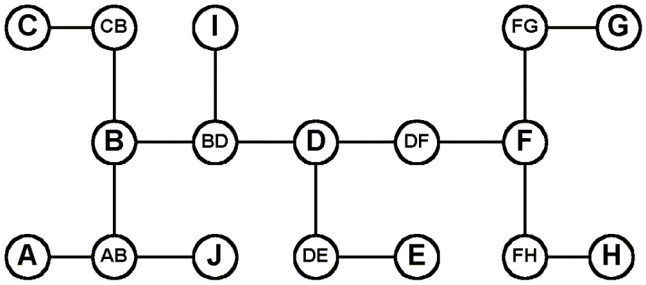

Fig. (4). Topological representation of environment.

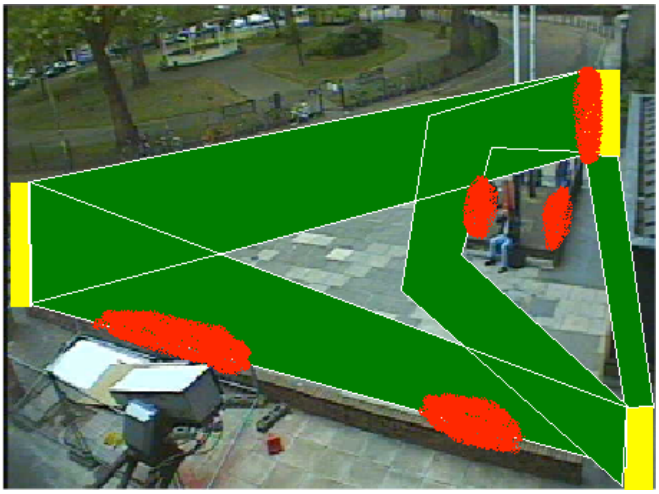

Fig. (5). Manually defined semantic labelling of a real surveillance scene. Entry/exit zones are shown by yellow rectangles, routes by green polygons and stop zones by red ellipses.

(EM) algorithm [18]. Detected entry/exit zones on the image planes of six different camera views are shown in Fig. (**7**). Stop events are also detected and localised where a target's speed falls below a threshold value. Similarly, stop zones can be modelled by GMM and learnt by EM.

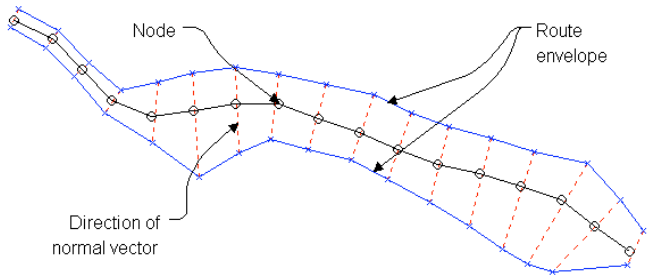

Fig. (6). Spatial-based route model. The central spline represents the mean average of the route and the side splines envelope the matching trajectories.

Routes are associated with the continuous "event" of a target's motion, which is described by a time sequence of location points (trajectory). We use a spline-based route model (Fig. **6**) to capture both the spatial extent and the usage of the routes. We have utilised the route model generated by an unsupervised learning method [17]. Fig. (**1**) illustrates the detected routes on the image plane of six different cameras.

Computational geometry can be used for further analysis of the scene routes that results in their deconstruction into primitives like paths and junctions. A junction is defined as the region of intersection of two routes where their directions differ by more than an angle ω. Paths are considered as route parts in between junctions and/or entry/exit zones. Because two routes may contain a common path (e.g., in Fig. (**3**), routes EDFG and HFG contain the path FG), the union of overlapped parts of routes with similar direction is also considered as a path.

The range of scene features that can be estimated is not restricted to the proposed scene model. Another example of the proposed scene reconstruction scheme is the estimation of occlusion areas on the image from occlusion events. An occlusion event can implicitly be detected by the motion tracking algorithm. If a blob has previously been successfully detected but it fails to be matched at the current frame, then its position is predicted, using Kalman filter. If the predicted position is within the scene, then the target is possibly occluded.

Fig. (**10**) depicts a scene where a synthetic occlusion (tape on the window) was imposed. A histogram $\mathbf{H_p}$ of occlusion events is given in Fig. (**11**) for a 10-hour period. A high rate of occlusion is evident in the heavily used area of the scene, mainly because of dynamic occlusions among vehicles and pedestrians. Although this histogram $\mathbf{H_p}$ does provide an estimate of the rate of occlusions, the interesting question is not "how many objects are occluded" but "how likely it is for an object to be occluded".

Therefore, it is necessary to take into account the activity in the area. Fig. (**12**) shows a histogram $\mathbf{H_m}$ of successfully matched blob positions over the same 10-hour period. Occlusions are represented by a probabilistic function H_o, defined by the following formula:

$$\mathbf{H}_{\mathbf{o}(i,j)} = \frac{\mathbf{H}_{\mathbf{p}(i,j)}}{\mathbf{H}_{\mathbf{m}(i,j)}} \qquad (1)$$

where i,j are indices to the pixels of the image. $\mathbf{H_o}$ is shown in the Fig. (**13**).

Although the current implementation allows identification of a limited range of spatial features, the principle that spatial features are related to specific events can provide a much wider range of semantics. For example a number of "turn left" events may be associated with a "turn-left" lane on a motorway, or a composite event "stop-queue-turn back" with accessing an ATM machine.

If events are related to classes of targets, then semantic labels can be more meaningful. For instance, pavements refer to paths used by pedestrians, or pedestrian crossings are junctions between a route used by a pedestrian and a route used by vehicles. Similarly, a bus stop can be related to the event sequence: "pedestrian stops-large vehicle stops-pedestrian merges with large vehicle".

Most surveillance systems contain multiple cameras and integration of the knowledge bases of different cameras is

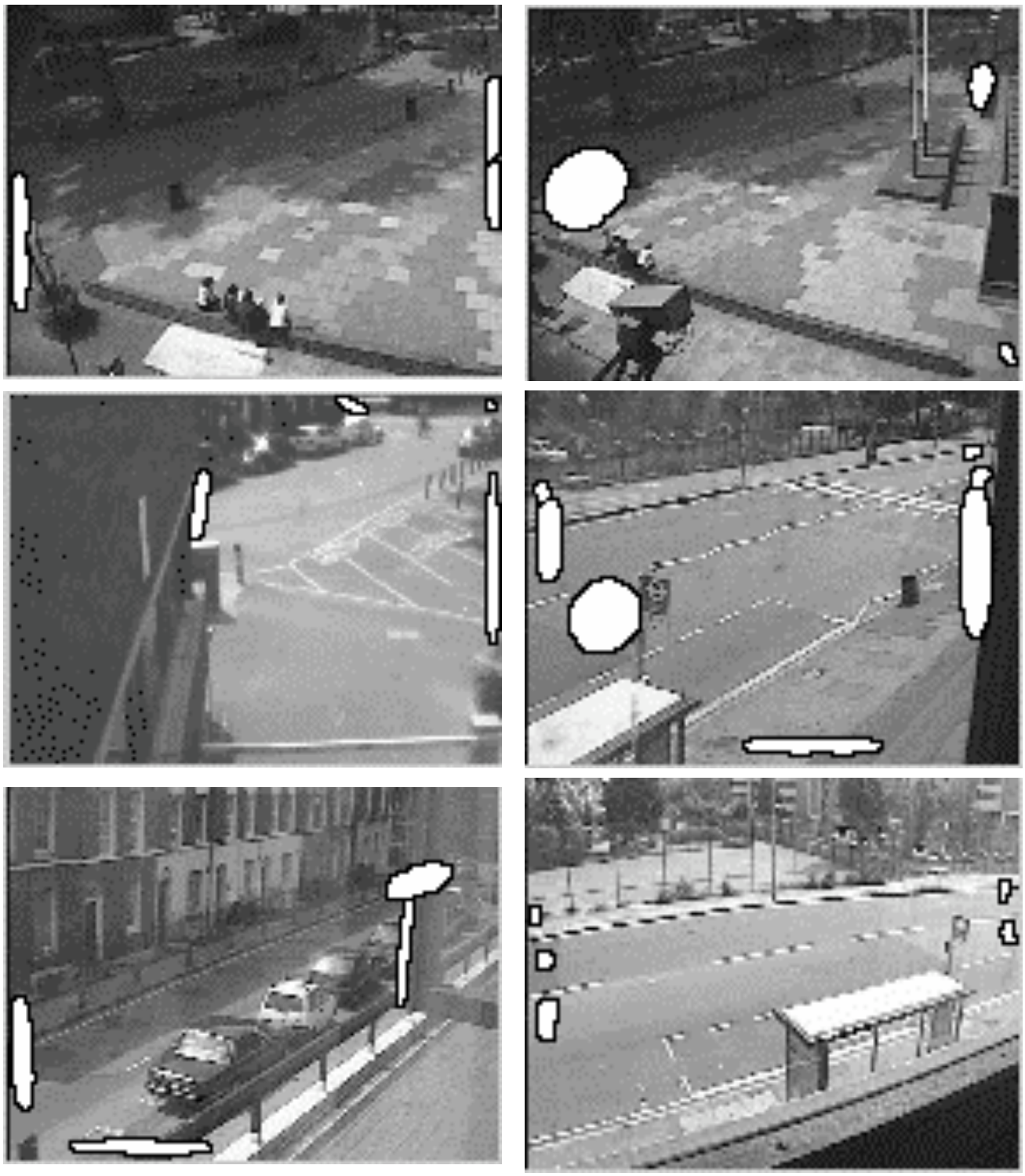

Fig. (7). Automatically derived entry/exit zones for the scenes of six camera views, learnt by sets of 16834, 19970, 13598, 6120, 10044, 30544 points respectively, are visualised by ellipses.

required. The traditional approach is to manually calibrate all the cameras, with respect to a common ground plane coordinate system. In this case, scene models can be learnt on the common ground plane. For instance, Fig. (**14**) illustrates the routes, detected in the 6 camera views of Fig. (**7**), projected on the common ground plane.

However the establishment of the common ground plane usually requires manual calibration that is a tedious task and if the camera is subsequently moved, the process of calibration must be repeated. If two camera views are substantially overlapped, a homography model of the two views can be automatically learnt [16] and is equivalent to the ground plane.

Fig. (8). Automatically derived routes for a set of six camera views, learnt by sets of 256, 500, 44, 500, 298 and 500 trajectories respectively.

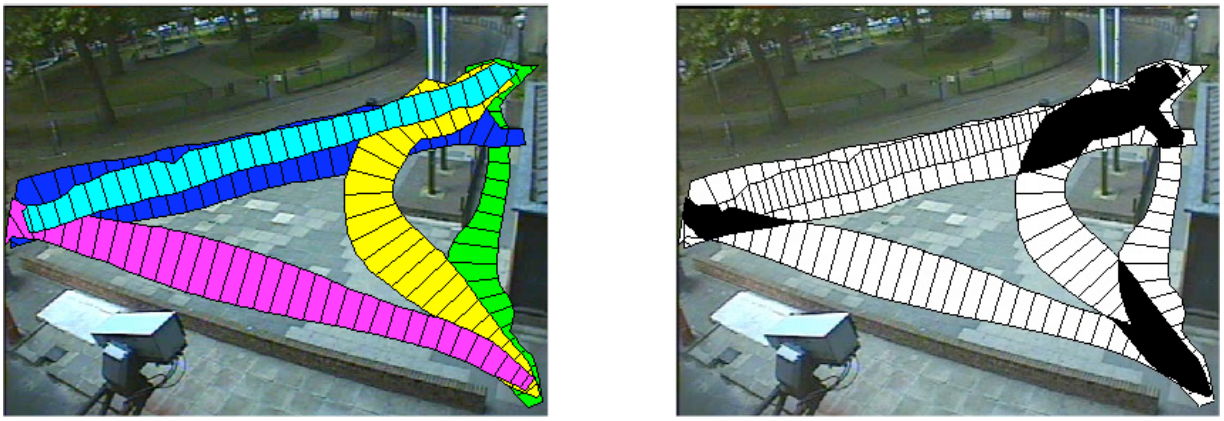

Fig. (9). Segmentation of a set of routes (left image) into junctions (black) and paths (white) (right image).

Fig. (10). A synthetic occlusion was produced by sticking some black tape on the window in front of the camera.

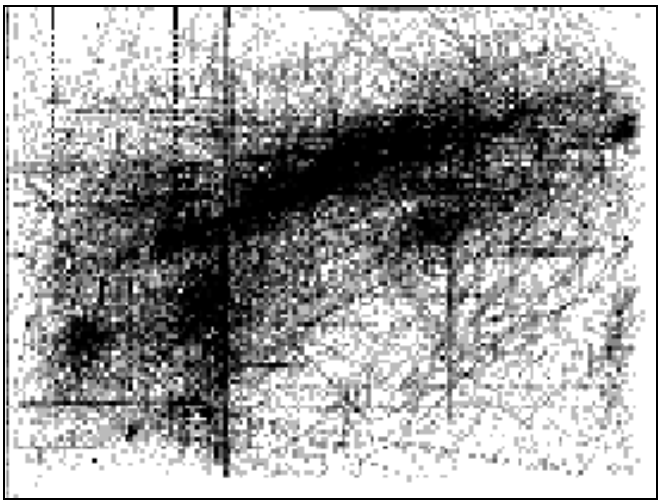

Fig. (11). Log histogram of 53614 occlusion events.

Fig. (12). Log histogram of 309380 matched positions.

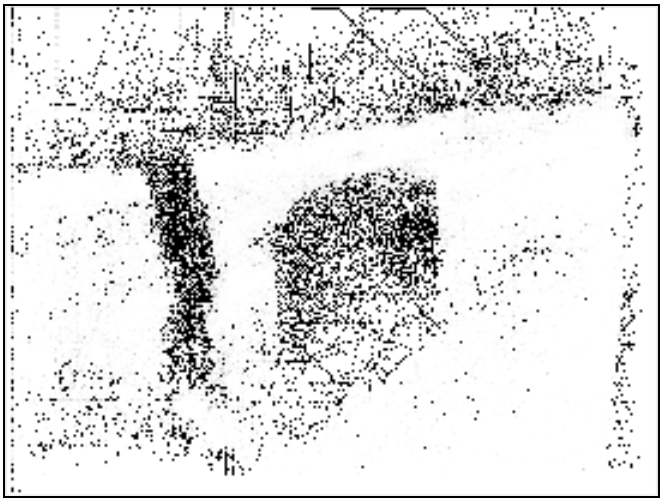

Fig. (13). Representation of H_o. The black tape areas and the bus stop sign are associated with large values of H_o.

We have developed a correspondence-free method [24] that can automatically learn the connectivity of a network of cameras (i.e: determine whether camera views are overlapped, adjacent or distant) and can build an integrated activity model for the entire observed scene. The method is based on the temporal correlation of entry/exit events between cameras and relates the different camera views in terms of transition times and transition probabilities. The method is able to relate adjacent cameras, even when their views do not overlap. Also, it provides a clue for the existence of "invisible" paths that are located in the gaps between camera views.

5. AUTOMATIC DETECTION OF SUSPICIOUS ACTIVITY

The automatic detection of suspicious activity is a major requirement for automated surveillance, because of the inability of human personnel to maintain their attention and monitor for suspicious activity over long periods of time. Our approach to the problem is similar in spirit with [24]. Specifically, we avoid modelling suspicious trajectories due to the very large variation of suspicious activity. Instead, we model normal trajectories and learn a probabilistic model of typical activity. Then, we associate suspicious behaviour with the atypical activity, i.e. the outliers of the typical activity.

Because the scene model is constructed from observations of the activity, it is related to patterns of typical activity. We have developed a method for automatic detection of atypical activity that may be suspicious, based on Route Based Hidden Markov Models (RBHMM), i.e. HMM overlaid on the routes of the scene model.

The states of a RBHMM are defined to be the nodes of all the accepted route models, plus two extra states: an “out-of-any-node state”, which indicates activity outside the modelled routes and an “end state”, which indicates the end of the observation. It is sensible to derive the nodes from unidirectional routes, so that directionality information is incorporated at each node.

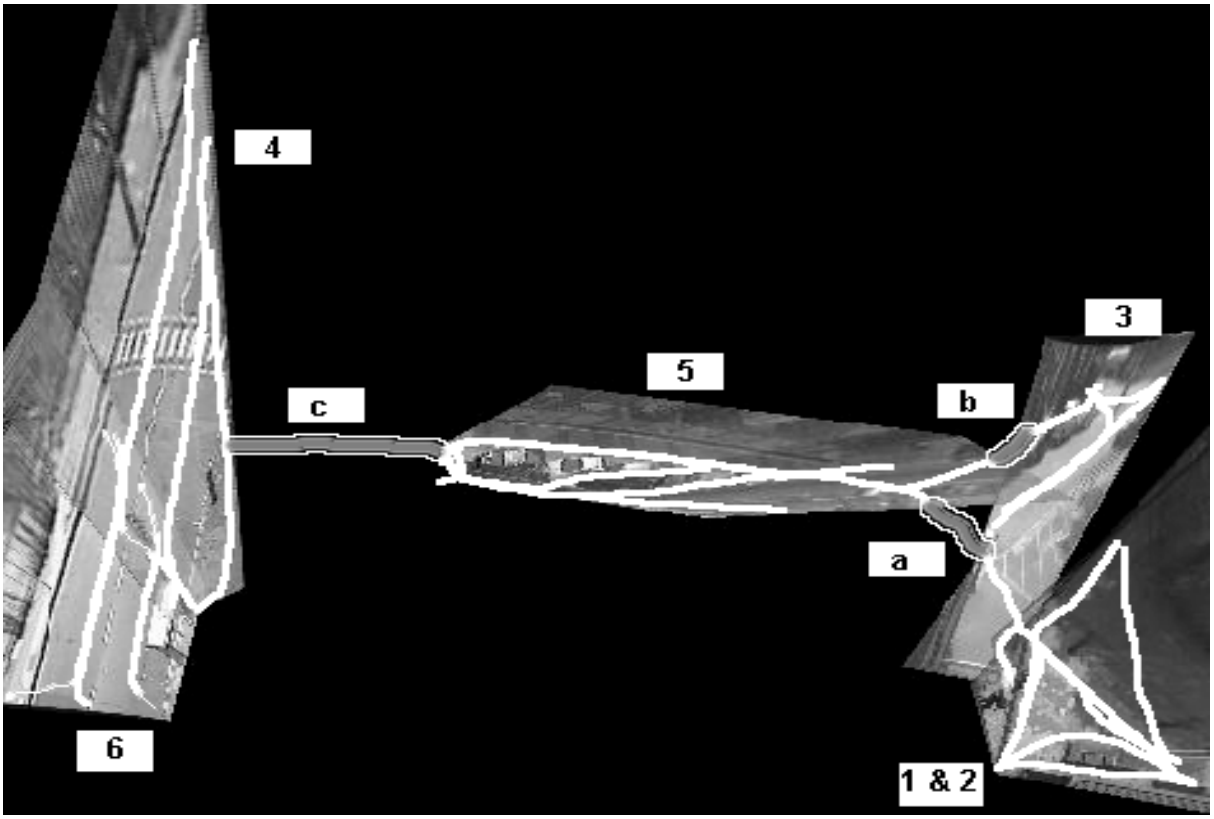

Fig. (14). Routes projected on the common ground plane of the six camera views. They consist of "visible" paths (in camera views 1-6) that were learnt from trajectories and "invisible" paths (in gaps a-c) that were determined by correlating entry/exit events.

Let assume that a scene contains W route models and each route model w=1..W consists of L_w nodes. The number of states N of the RBHMM is given by the following formula:

$$N = 2 + \sum_{w=1}^{W} L_w \tag{2}$$

The elements of the RBHMM are:

$S=\{S_i\}$, i=1..N, the set of states.

$Q=\{q_k\}$, k=1..M, the sequence of the states.

$A=\{a_{ij}\}$, i,j=1..N, the transition probability distribution, where $a_{ij}=P(q_t+1=S_j \mid q_t=S_i)$.

$\pi=\{\pi_i,\}$ i=1..N, the initial state distribution, where $\pi_i=P(q_1=S_j)$.

$O=\{O_k\}$, k=1..M, the sequence of the observations.

$B=[b_i(v)]$, i=1..N, v a position vector and $b_i(v)$ is the membership function of the observation v to the state i.

For such models, the HMM parameters are generally recommended to be learnt using iterative algorithms [10]. However, because of the large number of states, these algorithms are very slow and often impractical, especially for online learning. Instead, we use the pdf distributions of observations across the routes (see [18]) to encode the observation vector B (3). Then the RBHMM parameters are estimated cumulatively using (4) and (5):

$$b_i(O_{l.k}) = \frac{g_i(O_{l.k})}{\sum_{j}^{N} g_j(O_{l.k})} \tag{3}$$

$$\pi_i = \frac{\sum_{l=1}^{L} b_i(O_{l,1})}{L} \tag{4}$$

$$a_{ij} = \frac{\left(\sum_{l=1}^{L} \sum_{k=1}^{K_L} b_i(O_{l,k}) \cdot b_j(O_{l,k+1}) \right)}{\sum_{l=1}^{L} \sum_{k=1}^{K_L} b_i(O_{l,k})} \tag{5}$$

where $O_{l,k}$ is the kth observation of the lth trajectory l=1..L, k=1..K_L and $g_i(O_{l,k})$ is the estimate of the probability that the observation $O_{l,k}$ corresponds to the state i.

Trajectories can be evaluated according to the HMM and characterised as typical or atypical. Usually, suspicious events correspond to atypical trajectories, for instance the right image of the Fig. (**17**) that depicts the trajectory of someone appearing to climb up the wall.

More specifically, the consistency of an observation vector O (trajectory) with a given HMM λ, is represented by the probability $P(O|\lambda)$. The estimation of the above probability is known as the evaluation problem, the first of the three basic problems in HMM theory. The solution, as described in [10], is given by:

$$P(\mathbf{O} \mid \lambda) = \sum_{i=1}^{M} \alpha_M(i) \tag{6}$$

where M is the size of the observation vector O and αM is the forward variable used in the Viterbi algorithm. The typicality criterion for the observation vector is given as:

$$l(\mathbf{O}) = \log\left(P(\mathbf{O}|\lambda)\right)/M \tag{7}$$

The typicality criterion l(O) is not related directly to the probability P(O|λ), but to its logarithm, because the probability P(O|λ), may have too low a value to be represented within the arithmetic range of the computer. Division by M is performed to normalise the criterion against the size of the vector. Atypicality is detected when l(O) is below a threshold.

While the criterion defined by (7) indicates the typicality of an entire trajectory, the criterion defined in (8) characterises a specific sample O_k of the trajectory:

$$l'(\mathbf{O}_k) = \log\left\{P\left(O_1 O_2 \ldots O_k|\lambda\right)\right\} - \log\left\{P\left(O_1 O_2 \ldots O_{k-1}|\lambda\right)\right\} \tag{8}$$

or equivalently:

$$l'(\mathbf{O}_k) = \log\left(\sum_{i=1}^{k} \alpha_k(i) - \sum_{i=1}^{k-1} \alpha_{k-1}(i)\right) \tag{9}$$

Fig. (**15**) depicts three trajectories and Fig. (**16**) and Fig. (**17**) present their evaluation according to the two criteria. The first trajectory is a common trajectory that is verified by the values of the two criteria. The second trajectory is not so unusual, however it contains two samples (crosses) where the target speed increases. The criterion l estimates the overall typicality of the trajectory; therefore it does not find anything atypical. The criterion l' is able to detect the two suspicious samples. The third trajectory is a clearly an atypical trajectory that could represent somebody climbing (cross). The suspicious activity of the event is detected by both criteria.

6. DISCUSSION

The KUES Intelligent Visual Surveillance System consists of a variety of vision modules that operate at different levels. Additionally, it is required to operate in different environments and for extended periods and to "understand" the scene activity and alert the human operator. Therefore, they require being equipped with significant cognitive capabilities and it provided an opportunity to study the development of cognitive vision systems

We have presented a framework for cognitive surveillance systems that is able to derive a high-level understanding of activities. Description of activities is based on three main categories of semantics (targets, actions, events). Particularly, we focus on how a computer vision system can automatically learn semantic labels for fixed spatially-related entities in the scene, using a motion attention mechanism and exploiting the long-term consistencies of the visual data. We have built a scene model that includes event-based semantics that are learnt automatically from video data.

Fig. (15). Three trajectories are shown. The first trajectory (top) is very common. The second (middle) contains two rather suspicious samples (crosses). The third one (bottom) is a very uncommon one of somebody apparently climbing.

Automatic learning can also be used to create models for targets (e.g: separate pedestrians, vehicles and large vehicles) and to automatically identify the types of interesting events that take place in a particular scene. Such an approach will

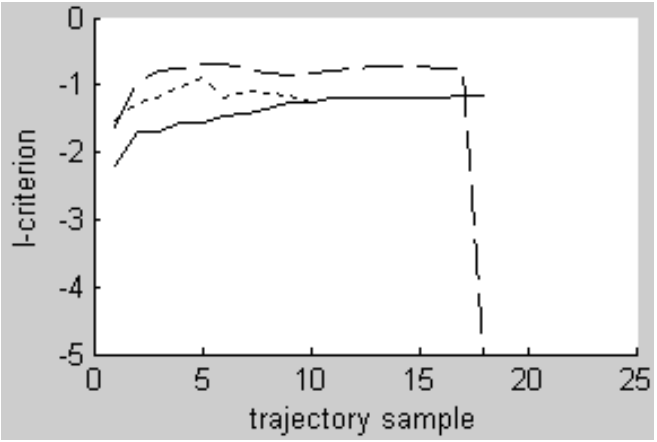

Fig. (16). The evaluation of the three trajectories according to the *l* criterion. The first (solid line) and the second (dotted line) trajectories of the Fig. (**15**) are in general typical, while the third one (dashed line) is clearly atypical.

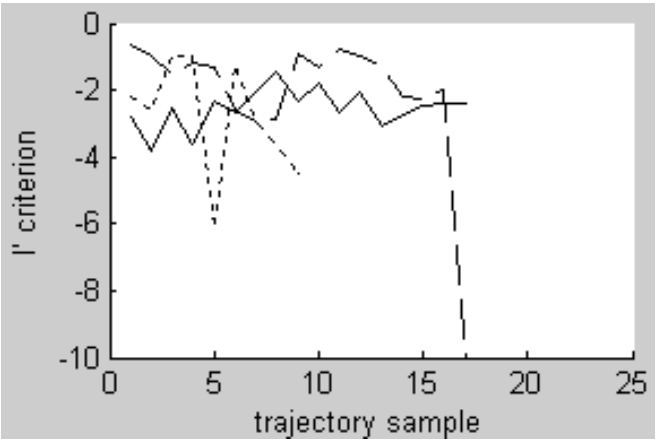

Fig. (17). The three trajectories of Fig. (**15**) are evaluated according to the l' criterion. All the samples of the first trajectory (solid line) are typical; two of the samples of the second trajectory (dotted line) are characterised as atypical; the climbing sample of the right (dashed line) is clearly atypical.

not only expand the range of semantics that can be automatically recognised by the surveillance system, covering all three categories of a surveillance dictionary, but also provide the basis of a system that is capable of distinguishing a wider variety of spatially-related semantics.

However, there are many open issues till the achievement of cognitive vision systems, such as the specification of appropriate knowledge representations and a mechanism to determine the interesting "semantics" for each scene.

ACKNOWLEDGEMENTS

The authors would like to thank Ming Xu for supplying the trajectory data. This work was in part supported by the EPSRC under grant number GR/M58030.

REFERENCES

[1] P. L. Rosin, "Thresholding for Change Detection", British Machine Vision Conference, BMVC97, 1997, pp. 212-221, Colchester, UK.

[2] C. Stauffer, W. E. L Grimson, "Adaptive background mixture models for real-time tracking". International Conference on Computer Vision and Pattern Recognition, CVPR99, Fort Colins, USA, 1999.

[3] S. S. Intille, J. W. Davis and A. F. Bobick, "Real-time closed-world tracking", International Conference on Computer Vision and Pattern Recognition, CVPR'97, 1997

[4] R. Rosales and S. Sclaroff, "Improved tracking of multiple humans with trajectory prediction and occlusion modelling", IEEE CVPR workshop on the Interpretation of Visual Motion, Santa Barbara, 1998.

[5] M. Isard, A. Blake, "Contour tracking by stochastic propagation of conditional density", European Conference on Computer Vision, ECCV96, vol.1, 1996, pp. 343-356, Cambridge UK.

A. F. Bobick, "Movement, Activity and Action: The Role of Knowledge in the Perception of Motion", Royal Society Workshop on Knowledge-based Vision in Man and Machine, London, UK, February 1997.

[6] F. Bobick, Y. A. Ivanov, "Action Recognition using Probabilistic Parsing", IEEE Conf. on Computer Vision and Pattern Recognition, CVPR98, Santa Barbara, CA, June 1998, pp.196-202.

[7] M. Brand, N. Olivier, A. Pentland, "Coupled Hidden Markov Models for Complex Action Recognition", IEEE Conference on Computer Vision and Pattern Recognition, CVPR97, Puerto Rico, 1997.

[8] N. J. Galata, D. C. Hogg, "Learning Variable Length Markov Models of behaviour", *Comput. Vis. Image Underst.,* vol. 81, pp. 398-413, 2001.

[9] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE, February. 1989, vol. 77, no. 2, pp. 257-286.

[10] Katz, J. Lin, C. Stauffer, E. Grimson, "Answering Questions about Moving Objects in Surveillance Videos", Proceedings of 2003 AAAI Spring Symposium on New Directions in Question Answering, 2003.

[11] H. Buxton, "Generative Models for Learning and Understanding Dynamic Scene Activity", Generative Model-Based Vision Workshop (GMBV2002) in ECCV 2002, Copenhagen, Denmark.

[12] M. Cohn, A. Galata, D. Hogg, S. Hazarika, "Towards an Architecture for Cognitive Vision using Qualative Spatio-Temporal Representations and Abduction", Proceedings of Spatial Cognition, June 2002.

[13] J. H. Fernyhough; A.G. Cohn, D. C. Hogg, "Generation of semantic regions from image sequences" in: Buxton, B & Cipolla, R (editors) Computer Vision ECCV'96, Springer-Verlag. 1996, pp. 475-478.

[14] J. Lou, Q. Liu, W. Hu, T. Tan, "Semantic Interpretation of Object Activities in a Surveillance System", International Conference on Pattern Recognition ICPR2002, Quebec, Canada, 2002.

[15] T. J. Ellis, J. Black, M. Xu, D. Makris, "A Distributed Multicamera Surveillance System", in Ambient Intelligence, a Novel Paradigm, Springer, 2005, pp. 107-138.

[16] Makris, T. J. Ellis, "Path Detection in Video Surveillance" in *'Image Vision Comput*., vol. 20, no. 12, pp. 895-903, October 2002.

[17] Makris, T. J. Ellis, "Learning Semantic Scene Models from Observing Activity in Visual Surveillance" in *'IEEE Trans. Sys. Man Cybern. Part B'*, vol. 35, no. 3, pp. 397-408, June 2005.

[18] A. F. Bobick, and J. W. Davis, "Recognition of Human Movement Using Temporal Templates", *IEEE Trans Pattern Anal. Mach. Intell.*, vol. 23, no. 3, 257-267, 2001.

[19] M. S. Yilmaz, "Recognizing Human Actions in Videos Acquired by Uncalibrated Moving Cameras", IEEE International Conference in Computer Vision (ICCV2005), Beijing, Chine, October 2005.

[20] J. Black, D. Makris, T. J. Ellis, "Hierarchical Database for a Multi-Camera Surveillance System" in 'Pattern Analysis and Applications', December 2004. vol. 7, no. 4, Springer, pp. 430-446.

[21] André, G. Herzog, T. Rist, "Natural Language Access to Visual Data: Dealing with Space and Movement", 1st Workshop on Logical Semantics of Time, Space and Movement in Natural Language, Toulouse, France, 1989.

[22] R. J. Howarth, H. Buxton, "Analogical Representation of Spatial Events for Understanding Traffic Behaviour". ECAI, pp.785-789. 1992.

[23] D. Makris, T. Ellis, J. Black, "Bridging the Gaps between Cameras", IEEE Conference on Computer Vision and Pattern Recognition, CVPR2004, Washington DC, USA, June 2004.

[24] N. Johnson. "Learning Object Behaviour Models", PhD thesis, School of Computer Studies, University of Leeds, UK, September 1998.

Received: March 31, 2008 Revised: May 30, 2008 Accepted: June 30, 2008