

Issues in Automated Visual Surveillance

Anthony R. Dick and Michael J. Brooks

School of Computer Science, University of Adelaide, Adelaide, SA 5005, Australia
CRC for Sensor, Signal and Information Processing, Technology Park, Mawson Lakes, SA 5095
{ard,mjb}@cs.adelaide.edu.au

Abstract. The usefulness of networks of surveillance cameras is primarily limited by the demand placed on human supervisors to monitor many real time video feeds simultaneously. The goal of automated visual surveillance is to reduce the burden on operators by including software in a surveillance system that can analyse behaviour automatically. This paper reviews progress in the field and considers some of the major remaining problems in automated video surveillance.

1 Introduction

Automated visual surveillance is currently a hot topic in computer vision research, and with good reason. To quote New Scientist magazine:

*If the technology takes off it could put an end to a longstanding problem that has dogged CCTV almost from the beginning. It is simple: there are too many cameras and too few pairs of eyes to keep track of them. With more than a million CCTV cameras in the UK alone, they are becoming increasingly difficult to manage.*¹

Surveillance cameras are cheap and ubiquitous, but the manpower required to supervise them is expensive. Consequently the video from these cameras is usually monitored sparingly or not at all; in fact it is often used merely as a record to examine an incident once it is known to have taken place. Surveillance cameras are a far more useful tool if instead of passively recording footage they can detect events requiring attention as they happen, and take action (for example alert a human supervisor) in real time. This is the goal of automatic visual surveillance: to obtain a description of what is happening in a monitored area, and then to take appropriate action based on video footage.

The nature of this description may vary according to the sort of decisions and actions a surveillance system is to make. For instance, a recent survey [45] indicated that high priorities for a public transport surveillance system are to detect congestion in restricted areas, and “individual delinquency” (e.g. violence against oneself or others). Detecting congestion requires only a simple description of each person in a scene, maybe just enough to count the number of people present, whereas the detection of delinquent behaviour requires a much richer description of an individual, possibly including a history of their overall motion, limb movements and gaze direction.

¹ New Scientist, 12 July 2003, page 4.

Although the exact requirements vary between surveillance systems, there are issues that are common to all (see also surveillance reviews in [8, 40]). Usually, an operator is interested only in certain objects in the scene. For instance, in surveillance of a public area, one may be interested only in monitoring the people within it rather than the entire scene. Isolating each person in each frame, and tracking them over time, is a problem of object detection, classification and tracking.² Once these objects of interest have been identified, a subsequent problem is that of behaviour analysis: describing their activity in such a way that a course of action can be decided upon in software.

Object identification and tracking, and behaviour analysis, are core problems of automatic surveillance, but they are affected by a number of practical problems. Surveillance networks contain many cameras, which need to be coordinated in order to track an object over an extended distance and period of time. The video obtained from these cameras is often of low resolution and frame rate, and varies in quality as environmental conditions such as lighting change over time. As well as arriving from many cameras, video may come from various types of camera, including infra-red and night-vision, as in the MIT Forest of Sensors project (part of VSAM [9]). To be useful, surveillance software usually needs to make decisions in real time, which further constrains the amount of processing that can be applied. It is important to take each of these issues into account when designing a surveillance system.

2 Principal challenges

Rather than advocate a set of “best practices” for the main problems in surveillance software, this section presents a range of approaches to each issue. This is for two reasons: first, the problems of object detection, tracking, identification and analysis from video are still open; and second, solutions to these problems tend to be highly domain specific.

An indication of the difficulty of creating a single general purpose surveillance system comes from the development of one of the most ambitious surveillance projects: the VSAM (Video Surveillance and Monitoring) project at CMU and other institutions [9]. VSAM is intended for automated surveillance of people and vehicles in cluttered environments, using a range of sensors including colour CCD cameras, thermal cameras and night vision cameras. This was intended as a general surveillance system, but has instead become a collection of separate algorithms which are selected on a case by case basis (many of which, it should be noted, advance the state of the art in surveillance research).

2.1 Object detection and tracking

Various techniques have been advocated for detecting and tracking objects in video. Corner and edge features can be clustered together to form objects, and then tracked [3]. Alternatively “snake” contours can be used to detect an object outline and then track

² This review is not concerned with biometrics, which involves the identification of individuals using face recognition or some other means. Rather, it focuses on techniques for detecting kinds of objects which are likely to be of interest, such as people, or cars or tanks.

it across frames [23, 18]. Pixel or region based background subtraction techniques (reviews appear in [21, 31]), in which a model of foreground and/or background appearance is learnt, blob trackers (e.g. [47]) and variations on optical flow (e.g. [16, 11]) have also been used. A Bayesian approach to region based object detection and tracking without background subtraction is presented in [44].

There is in general a tradeoff in the choice of approaches to detection and tracking between primitives that are easy to detect, but difficult to track consistently (such as corner features) and those that are easier to track, but difficult to detect, such as higher level shape models. Surveillance generally demands that objects are tracked over long periods of time, and in varying conditions. This raises difficulties such as tracking in very different lighting conditions (possibly day and night), across a cluttered and dynamic background, and in the presence of shadows.

Because tracking in surveillance video is difficult, an *a priori* model of the target (such as an articulated human body model, or simple model of a car parameterised on width and height [12]) is often used. For example, W4 (Who? When? Where? What?) [15] is designed for tracking people using a combination of learnt textural appearance models and contour tracking. The “cardboard model” of human shape is quite strong and can therefore be used to track people across cluttered backgrounds, through partial occlusion and in groups, and to detect if a person has picked up an object. It is also important that trackers are able to adapt over time to varying conditions, for instance by including temporal decay or multi-modal feature distributions [43] in their object model.

The need to automatically detect suspicious objects, such as a suitcase left in an airport lounge, has led to the development of systems for detecting objects that come to rest for a period of time. This has been solved in [13] by regularly computing a measure of the background exhibited in a video sequence of a busy scene using histogram methods, and detecting whether changes to the background have occurred that indicate a new static object has entered the scene. Such a system has obvious additional application in the detecting of blocked road tunnels and illegally parked vehicles.

2.2 Object classification

Object recognition is a classic computer vision problem, which has been tackled in a variety of ways (one review appears in [38]). Surveillance footage usually has quite poor resolution, and objects of interest may span only a few pixels in each frame. This lack of information means that generally coarse colour histogram techniques (e.g. [39]) are most applicable on a frame by frame basis. On the other hand, footage is available over a long period of time, which enables an informative model of motion to be constructed [10, 14].

For example, VSAM uses two classification algorithms, both of which require training. The first algorithm is a neural network which is trained on blob shape and area, and can discriminate individual humans, human groups, vehicles and clutter. The second is LDA (Linear Discriminant Analysis) performed on 11 dimensional feature vectors which include blob position, width, height and image features within the blob. Both algorithms are reported to have approximately a 90% success rate, although LDA appears

to be able to discriminate slightly more finely than the neural net (for instance, discriminating between cars and trucks) because it incorporates more features. Both classifiers operate on single frames, but results from previous frames are cached for smoothing.

There is considerable military interest in the analysis of video surveillance; see the DARPA Airborne Video Surveillance project,³ for example. Work done for this project includes the detection and classification of objects of interest such as people and vehicles, based on the periodicity of their motion [10]. The system is reportedly able to differentiate bipedal (people), quadrupedal (dogs) and “other” objects from some aerial footage.

3 Behaviour analysis

Object detection, tracking and classification, though unsolved problems in themselves, can be seen as precursors to the main problem in automated surveillance: the description of the activity taking place in the scene, which is usually termed behaviour analysis.

3.1 Human motion analysis

One of the most popular and demanding types of behaviour to analyse automatically is that of a human being [1]. A common approach to describing human motion is to use a state-based model, such as a Hidden Markov Model (HMM), to convert a series of motions into a description of activity. Such systems [6, 42, 19, 29, 35, 32] operate by training a HMM (or some variant thereon) to parse a stream of short-term tracked motions, analogous to the way speech recognition works by parsing a stream of phonemes. Each system has slightly different capabilities: for example [35] is able to classify interactions between a pair of people, such as changing direction to approach one another, talking together, and parting, on a helpful background (chequerboard floor, fairly barren backdrop). Another system [19] recognises simple human gestures (against a black backdrop), while an earlier system [42] recognises simple actions (pick up, put down, push, pull, etc.), also based on a trained HMM. The detection of anomalous behaviour is addressed in [32] for a security camera surveying an office corridor, to try to detect loitering or forced entry to an office.

A different approach is taken by Wada et al. [46] who use a “hypothesise and test” algorithm to interpret and predict human behaviour in a pseudo-office environment (with white markers placed on a black floor). Hypotheses are generated from a classification network (similar to a HMM, but admitting multiple solutions) which is trained in the same environment in which it is used.

A finite state machine can also be used to recognise a limited set of human behaviours [2]. This system relies on a great deal of prior knowledge about the layout of the office environment in which it is used, and the order in which actions can occur, which defines the structure of the state machine. It is less flexible than a Hidden Markov model, as all possible event transitions are explicitly modelled before any data is seen, but requires no training.

³ <http://www.darpa.mil/SPO/programs/airbornevideosurveillance.htm>

Rather than using a HMM to learn and then recognise human activity, Yacoob and Black [48] use a template based approach. Initially, measurements of canonical motions (for instance, one stride of a person walking) are recorded as exemplars of each activity they want to recognise. These measurements are stacked into a matrix on which principle component analysis is performed to extract dominant feature vectors for each activity. Activity is then recognised in subsequent footage by minimising a simple sum of squared distance measure. This algorithm is demonstrated recognising 4 different types of walk, using about 15 features, with about an 80% success rate.

Mann et al. [30] analyse video content by physically modelling the interaction of objects in video. However this requires the complete specification of the geometry of each object in the video, and is limited to very simple interactions such as picking up a can of soft drink or tipping one box against another.

In [27], the observed behaviour is reported in natural language sentences rather than as a set of disjointed phrases. Behaviour is recognised, and its natural language description built, in a bottom up hierarchy (starting with the motion of body parts and building them into an overall activity description). As with other behaviour recognition systems, however, the range of recognised activities is small (it is claimed to have over 30 verbs in its vocabulary, but only 6 are demonstrated in a fairly bare room containing 9 well separated objects).

VSAM also includes a couple of algorithms for higher level activity analysis, although no results are presented for these. The first deduces a simple skeleton from a human blob outline and attempts to deduce whether the human is walking, running or standing still based on analysis of the skeleton over time. The second is a Markov model, although this is limited to a small number of interactions and has to be trained for those situations.

3.2 Traffic motion analysis

Although the behaviour and motion of traffic in an urban area is quite different in nature to human behaviour, similar techniques apply to its analysis.

Brand and Vettaker [4] present a system that learns patterns of behaviour by training a HMM. Anomalous behaviour, such as a car driving in the wrong lane or turning left from the right hand lane, is then detected from a video camera mounted above an intersection. This system is also applied in an office environment to detect unusual behaviour (such as falling asleep, or standing at the window, apparently).

The Bobick system [19] is demonstrated interpreting movement in a car park. The system can detect events such as a car entering or leaving the car park, a person entering or leaving the car park, a person being picked up or dropped off, or losing or finding a track.

A system at the University of Reading [12] is designed more specifically for the tracking and analysis of traffic from a static or vehicle mounted camera. It fits a 3D outline of each car to the video, tracks its position and velocity and predicts its likely future motion, raising an alarm when a collision is imminent (though there are no actual examples of this!). A similar system using tracked feature points is described in [5].

4 Practical considerations

The difficulty of the tasks of object detection, tracking and analysis is compounded by a number of practical problems. These include keeping track of objects as they move between camera fields of view, ensuring that video is delivered and processed in real time, dealing with video of varying quality, and evaluating the success or otherwise of a surveillance system.

4.1 Robustness

If a good laboratory solution to a key problem in surveillance is to be translated into a commercial system, a good deal of further work is likely to be necessary to achieve suitable robustness in performance [37]. For example, referring back to the problem of detecting objects posited in a busy airport lounge, a commercial system will need to have a very low false-negative rate (we rarely overlook a suspicious object), perhaps at the expense of a slightly elevated false positive rate (we sometimes mistakenly detect an object). This in turn will require that excellent system performance is maintained in the face of difficult and fluctuating operating conditions. Illumination variation presents particular difficulties in practice, especially in monitoring outdoor scenes. Shadows may need to be isolated and ignored, and illumination changes associated with moving clouds, the onset of rain, setting of the sun, etc., may need to be handled. (See [33] for an example of research on vision in bad weather.) Indoor lighting changes can also present considerable problems when there are windows, intermittent light-source occlusions, reflections, doors opened, saturation effects due to bright lights, etc. Commercial systems clearly need to be as immune as possible to these types of problems in so far as they arise in the domain of application. Additionally, systems are likely to need variable thresholding across the image for object detection sizes and characteristics. Such facilities, amongst others, have been built into the iOmniscient commercial system⁴ arising out of [13].

4.2 Camera handoff

A common assumption in multi-camera surveillance systems is that the fields of view (FOVs) of each camera overlap. This can be quite a severe restriction, especially when monitoring a wide area or one where lines of sight are often occluded. Camera handoff is a term commonly used to describe algorithms for keeping track of an object across multiple camera FOVs, in order to overcome this limitation.

Early camera handoff algorithms [24, 7] require both camera calibration and overlapping fields of view to maintain tracks across cameras. Additionally, a 3D model of the environment is required by [24]. Overlapping FOVs are required to register views relative to each other by tracking an object while it is visible in more than one view, while calibration data is used to determine the object's world position, and therefore its visibility in each camera. Subsequent systems relax the need for camera calibration

⁴ <http://iomniscient.com>

information [28, 20, 26, 22], but still require overlapping views to establish correspondences.

Kettnaker [25] presents a Bayesian approach to tracking objects through FOVs that do not overlap. However this system requires a set of allowable paths, and a set of transition probabilities and times, to be given as input.

The surveillance of traffic is well suited to cameras with non-overlapping FOVs, because traffic generally follows well defined paths and these paths extend over long distances which it is impractical to monitor entirely. A system using cameras situated about 2 miles apart along a highway is described in [17, 36]. Cars are identified as they enter each camera's field of view based on both their appearance and positional information from views through which they have previously passed.

4.3 Efficiency

Surveillance analysis software is usually more useful if it can run in real time, and is therefore able to act on what it sees as it is happening. Although increases in CPU and memory speed have enabled far more sophisticated techniques to be executed in real time, there remains a network bandwidth bottleneck: video streams from several sources need to be collected over a network and moved to where they can be accessed by a CPU. Care therefore needs to be taken to minimise network traffic within a surveillance system. For example, only frames in which some activity is present need to be transmitted and processed. This presumes some simple processing ability (e.g. motion detection) near each video source—these are called Sensor Processing Units in VSAM [9].

4.4 Evaluation of surveillance systems

Surveillance systems tend to be quite domain specific, meaning that the best test of the merit of a system is usually to test it extensively in the environment in which it is to be used. However this is not always possible, and some work has been done on quantitative testing procedures using Receiver Operating Characteristic (ROC) curves [34]. Such measures are also useful for testing a system with varying parameters on a single testbed, and obtaining results which can be meaningfully compared.

Surveillance is an area of vision which particularly benefits from a standard set of test data, and some progress has been made towards this goal, such as the PETS database⁵, which is associated with the annual PETS (Performance Evaluation of Tracking and Surveillance) workshop.

5 Surveillance video mining

In addition to real-time surveillance of events, there is an evident need for after-the-event analysis of stored video. Public cameras are now ubiquitous to the extent that a person going about their business in a typical large city can expect to be video-taped hundreds of times per day. As a consequence, when an event such as a robbery takes

⁵ <ftp://pets.rdg.ac.uk>

place, it is common for police to scan (by eye) numerous video tapes from suitably positioned CCTV systems in search of images of an offender. As video repositories switch increasingly from tape to digital form, this presents a variety of challenges for automated surveillance, some of which relate to the field of video mining [41]. For example, a system may be required to search a video repository's last few days' footage for instances of violent behaviour, or examples of where a car passed unusually quickly through the field of view.

If several thousand networked digital video repositories are associated with cameras distributed across a city or building complex, a major challenge would be to track a vehicle or person across the network. Thus, if a robbery takes place at a given location and video information is obtained from a repository associated with a nearby camera, can the escape route be determined by searching through the network of repositories? Clearly, if the spatial distribution of roads and cameras is represented, various intelligent strategies could be used to reduce the extent of the search. This may be regarded as a scaled up version of the camera handoff problem described in Section 4.2, and remains a tremendous challenge.

6 Discussion

Despite recent progress in computer vision and other areas, there are still major technical challenges to be overcome before the dream of reliable automated surveillance is realised. These technical challenges, including object identification, tracking and analysis, are compounded by practical considerations such as the physical placement of cameras, the network bandwidth required to support them, installation cost, privacy concerns and robustness to unfavourable weather and lighting conditions. However progress is being made ever more rapidly, and the demand for automated surveillance continues to increase in areas ranging from crime prevention, public safety and home security to industrial quality control and military intelligence gathering. To end where we began, with a quote from *New Scientist*:

In a survey of hundreds of US security executives, he found that systems which could process the video from the spiralling number of CCTV cameras were "one of the top items in demand".⁶

References

- [1] J.K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440, March 1999.
- [2] D. Ayers and M. Shah. Monitoring human behavior from video taken in an office environment. *Image and Vision Computing*, 19(12):833–846, October 2001.
- [3] P.A. Beardsley, A. Zisserman, and D.W. Murray. Sequential updating of projective and affine structure from motion. *International Journal of Computer Vision*, 23(3):235–259, 1997.

⁶ *New Scientist*, 12 July 2003, page 5.

- [4] M. Brand and V. Kettner. Discovery and segmentation of activities in video. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):844–851, August 2000.
- [5] P. Burlina and R. Chellappa. Temporal analysis of motion in video sequences through predictive operators. *International Journal of Computer Vision*, 28(2):175–192, 1998.
- [6] H. Buxton. Learning and understanding dynamic scene activity: a review. *Image and Vision Computing*, 21(1):125–136, January 2003.
- [7] Q. Cai and J. Aggarwal. Tracking human motion using multiple cameras. In *Proc. International Conference on Pattern Recognition*, pages 68–72, 1996.
- [8] R. Collins, A. Lipton, and T. Kanade. Introduction to the special section on video surveillance. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):745–746, August 2000.
- [9] R.T. Collins, A.J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, and O. Hasegawa. A system for video surveillance and monitoring. Technical Report CMU-RI-TR-00-12, CMU, 2000.
- [10] R. Cutler and L.S. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):781–796, August 2000.
- [11] L. S. Davis, S. Fejes, D. Harwood, Y. Yacoob, I. Haratoglu, and M. J. Black. Visual surveillance of human activity. In *Asian Conference on Computer Vision*, pages 267–274, 1998.
- [12] J.M. Ferryman, S.J. Maybank, and A.D. Worrall. Visual surveillance for moving vehicles. *International Journal of Computer Vision*, 37(2):187–197, June 2000.
- [13] D. Gibbins, G. Newsam, and M. J. Brooks. Detecting suspicious background changes in video surveillance of busy scenes. In *Third IEEE Workshop on Applications of Computer Vision*, pages 22–26, 1996.
- [14] R. Goldenberg, R. Kimmel, E. Rivlin, and M. Rudzsky. 'Dynamism of a Dog on a Leash' or Behavior classification by eigen-decomposition of periodic motions. In *Proc. 7th European Conference on Computer Vision*, page I: 461 ff., 2002.
- [15] I. Haritaoglu, D. Harwood, and L.S. Davis. W4: Real-time surveillance of people and their activities. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):809–830, August 2000.
- [16] B. K. P. Horn and B. G. Schunk. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.
- [17] T. Huang and S. Russell. Object identification in a Bayesian context. In *Proceedings of IJCAI*, pages 1276–1283, 1997.
- [18] M. Isard and A. Blake. Condensation—conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.
- [19] Y.A. Ivanov and A.F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):852–872, August 2000.
- [20] O. Javed, S. Khan, Z. Rasheed, and M. Shah. Camera handoff: Tracking in multiple uncalibrated stationary cameras. In *Workshop on Human Motion*, pages 113–118, 2000.
- [21] O. Javed, K. Shafique, and M. Shah. A hierarchical approach to robust background subtraction using color and gradient information. In *Workshop on Motion and Video Computing (MOTION'02)*, pages 22–27, 2002.
- [22] J. Kang, I. Cohen, and G. Medioni. Continuous tracking within and across camera streams. In *Proc. IEEE Computer Vision and Pattern Recognition*, pages 267–272, 2003.
- [23] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1:321–331, 1987.
- [24] P. Kelly, A. Katkere, D. Kuramura, S. Moezzi, and S. Chatterjee. An architecture for multiple perspective interactive video. In *Proceedings of the 3rd ACM International Conference on Multimedia*, pages 201–212, 1995.

- [25] V. Kettner and R. Zabih. Bayesian multi-camera surveillance. In *Proc. IEEE Computer Vision and Pattern Recognition*, pages 253–259, 1999.
- [26] S. Khan, O. Javed, Z. Rasheed, and M. Shah. Human tracking in multiple cameras. In *Proc. IEEE International Conference on Computer Vision*, pages I:331–336, 2001.
- [27] A. Kojima, T. Tamura, and K. Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 50(2):171–184, November 2002.
- [28] L. Lee, R. Romano, and G. Stein. Monitoring activities from multiple video streams: Establishing a common coordinate frame. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):758–767, August 2000.
- [29] J. Lou, Q. Liu, T. Tan, and W. Hu. Semantic interpretation of object activities in a surveillance system. In *Proc. International Conference on Pattern Recognition*, pages III: 777–780, 2002.
- [30] R. Mann, A.D. Jepson, and J.M. Siskind. The computational perception of scene dynamics. *Computer Vision and Image Understanding*, 65(2):113–128, February 1997.
- [31] A. McIvor, Q. Zang, and R. Klette. The background subtraction problem for video surveillance systems. Technical Report CITR-TR-78, University of Auckland, 2000.
- [32] V. Nair and J.J. Clark. Automated visual surveillance using hidden markov models. In *International Conference on Vision Interface*, pages 88–93, 2002.
- [33] S. G. Narasimhan and S. K. Nayar. Chromatic framework for vision in bad weather. In *Proc. IEEE Computer Vision and Pattern Recognition*, pages 1598–1605, 2000.
- [34] F. Oberti, E. Stringa, and G. Vernazza. Performance evaluation criterion for characterizing video-surveillance systems. *Real Time Imaging*, 7(5):457–471, October 2001.
- [35] N.M. Oliver, B. Rosario, and A.P. Pentland. A Bayesian computer vision system for modeling human interactions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):831–843, August 2000.
- [36] Hanna Pasula, Stuart J. Russell, Michael Ostland, and Yaacov Ritov. Tracking many objects with many sensors. In *Proceedings of IJCAI*, pages 1160–1171, 1999.
- [37] I. Pavlidis, V. Morellas, P. Tsiamyrtzia, and S. Harp. Urban surveillance systems: from the laboratory to the commercial world. *Proceedings of the IEEE*, 89(10):1478–1497, 2001.
- [38] A.R. Pope. Model-based object recognition: A survey of recent research. Technical Report 94-04, University of British Columbia, 1994.
- [39] Y. Raja, S. J. McKenna, and S. Gong. Tracking and segmenting people in varying lighting conditions using colour. In *IEEE International Conference on Face and Gesture Recognition*, pages 228–233, 1998.
- [40] C. Regazzoni, V. Ramesh, and G. Foresti. Scanning the issue/technology: Special issue on video communications, processing and understanding for third generation surveillance systems. *Proceedings of the IEEE*, 89(10):1355–1366, 2001.
- [41] A. Rosenfeld, D. Doermann, and D. DeMenthon (eds). *Video Mining*. Kluwer, 2003.
- [42] J.M. Siskind and Q. Morris. A maximum-likelihood approach to visual event classification. In *Proc. 4th European Conference on Computer Vision*, pages II:347–360, 1996.
- [43] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.
- [44] J. Sullivan, A. Blake, M. Isard, and J.P. MacCormick. Bayesian object localisation in images. *International Journal of Computer Vision*, 44(2):111–135, September 2001.
- [45] S. Velastin, M. Sanchez-Svensson, J. Sun, M. Vicencio-Silva, D. Aubert, A. Lemmer, P. Brice, L. Khoudor, and S. Kallweit. D7P: Innovative tools for security in transports. Technical Report GRD1-2000-10601, PRISMATICA Project, 5th Framework Programme, 2003.

- [46] T. Wada and T. Matsuyama. Multiobject behavior recognition by event driven selective attention method. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):873–867, August 2000.
- [47] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.
- [48] Y. Yacoob and M.J. Black. Parameterized modeling and recognition of activities. *Computer Vision and Image Understanding*, 73(2):232–247, February 1999.