Towards intelligent networked video surveillance for the detection of suspicious behaviours

M.J. Brooks, A.R. Dick, A. van den Hengel
School of Computer Science, University of Adelaide, SA 5005, AUSTRALIA

ABSTRACT: *Video camera surveillance is becoming increasingly pervasive in both outdoor city environments and indoor complexes. Indeed, it is reckoned that a central Londoner is captured on well over 300 videos on an average day. With the generation of this vast repository of visual data come the challenges of automated real-time detection of key events and retrieval of salient recorded information. The last 10 years has seen the development of a range of computerised technologies for automated surveillance. Technology of the authors for detecting objects left unattended in domains such as airports has resulted in successful commercialisation, worldwide sales and the winning of international awards. However, the need is now greater than ever for a surveillance system that is truly networked. Such a system will track individual people and vehicles through cluttered environments using hundreds of cameras. With this tracking information, the system will be able to detect a range of anomalous behaviours, such as the suspicious movement of a person, or the illegal manoeuvre of a vehicle. In this paper, we outline the state of the art in video surveillance and suggest that graph-based representations will prove valuable in taking surveillance to a truly networked level of operation.*

1 INTRODUCTION

In recent years, automated visual surveillance has become a much-researched topic in computer vision, 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.
>
> New Scientist, 12 July 2003, p.4

Surveillance cameras are cheap and ubiquitous, but the manpower required to supervise their output is expensive. Consequently the video from cameras is usually monitored sparingly or not at all. Indeed, video is most commonly used as an archive enabling referral back to an event, but only once an incident is known to have taken place. Even if adequate people are assigned to view a bank of monitors associated with video cameras, it is well known that human attention span and observancy decline rapidly over a short period.

Rather than simply being used in a passive recording mode, surveillance systems are far more useful when in real time they are able to automatically detect key events and take action (for example, alert a human supervisor). Thus, a prime goal of automated visual surveillance is to obtain a live description of what is happening in a monitored area and take (or trigger) appropriate action.

Examples of activities that a surveillance system might undertake are detections of a person:

- Entering a particular area.
- Removing an object.
- Leaving a bag unattended.
- Behaving violently or erratically.
- Writing graffiti.

Naturally, a system may alternatively oversee traffic activity (detecting illegal manoeuvres or parking) or the buildup of crowds (congestion), etc.

Not always appreciated is that visual tasks people find straightforward can sometimes represent major challenges for the computer. The computational effort and complexity involved in simply "following" someone through an extended video sequence is enormous, and a truly robust and reliable tracker has yet to be developed. Compounding the problem is that surveillance cameras usually exhibit relatively low resolution, public areas under surveillance often have fluctuating and variable lighting conditions, people are frequently occluded by other people or structures, and people may

temporarily leave a monitored area, etc. Each of these factors can add tremendous difficulty to the task. Nevertheless, major progress has been made in recent years towards the development of sophisticated automated surveillance systems, and some of this will be outlined below.

Ubiquitous monitoring of public, workplace and other areas inevitably highlights privacy concerns. Explicit video surveillance legislation across the Australian states, specifying what may be monitored and how information may be used, is typically non-existent or incomplete. Having said this, video-surveillance privacy concerns appear to have subsided over the last ten years. In a recent survey concerning the London Transport system, users stated that their prime desire was for intensive video surveillance that would render their travels safer by detection of "individual delinquency" (Velastin et al., 2003). Typically, people are content to have a shopping precinct car park monitored if it means that cars and shoppers are less vulnerable to crime. Having said this, there is evidence that monitoring of troublesome street areas acts largely to move crime elsewhere. Further discussion of privacy concerns is beyond the scope of this paper.

In the following sections, we consider the state of the art in video surveillance, the need for truly networked surveillance, and offer some suggestions as to how this might be achieved. Finally, some future prospects are discussed. Note that face recognition will not be discussed in this article as the subject is regarded as something of a separate, specialised field. As it happens, though, face recognising systems have yet to attain a level of performance that would render them valuable in standard surveillance environments.

2 STATE OF THE ART

Key tasks in video surveillance are object detection, identification, tracking, and analysis of behaviour. We now discuss the state of the art in these and other areas before describing an example of a successful commercial video surveillance system.

2.1 *Detection and Identification*

The level of sophistication required of a system for detecting and possibly identifying objects in video sequences depends on the target application. For example, some simple problems can be solved merely by detecting that a moving object has entered a given space. Detecting congestion similarly requires only a basic description of each person, perhaps just enough to count the number of people present in an area. Detection of delinquent behaviour, on the other hand, requires a much richer description of an individual, possibly including a history of their overall motion, limb movements and gaze direction. Building up this history requires tracking, or following an individual through a video sequence.

Various techniques have been advocated for detecting and tracking objects in video. Corner and edge features can be clustered together to form objects, and can then be tracked (Beardsley et al., 1997) Alternatively "snake" contours can be used to detect an object outline which is then tracked across frames (Isard & Blake, 1998). Pixel or region based background subtraction techniques (reviews appear in Javed et al., 2002, McIvor et al., 2000) in which a model of foreground and/or background appearance is learnt have also been used, as have blob trackers and variations on optical flow

Object identification is a classic computer vision problem that has been tackled in a variety of ways (one review appears in Pope, 1994). 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 are most applicable on a frame-by-frame basis (e.g. Raja et al., 1998). On the other hand, footage is available over a long period of time, which enables an informative model of motion to be constructed.

For example, the VSAM system uses two identification algorithms, both of which require training. The first algorithm is a neural network that is trained on blob shape and area, and can discriminate individual humans, human groups, vehicles and clutter. The second is an LDA (Linear Discriminant Analysis) method performed on 11-dimensional feature vectors that 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; for example, see the DARPA Airborne Video Surveillance project[1]. 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 (Cutler & Davis, 2000). The system is reportedly able to differentiate bipedal (people), quadrapedal (dogs) and "other" objects from aerial footage.

[1] http://www.darpa.mil/SPO/programs/airbornevideosurveillance.htm

2.2 *Tracking*

There is in general a tradeoff in object 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 (e.g. a 3D CAD model of the target). Surveillance generally demands that objects are tracked over long periods of time, and in varying conditions. This raises difficulties such as tracking in varying lighting conditions (possibly day and night), across a cluttered and dynamic background, and in the presence of shadows.

Because tracking in surveillance video can be so demanding, an a priori model of the target (such as an articulated human body model, or a simple model of a car parameterised on width and height) is often used. For example, the W4 system (W4 being short for Who? When? Where? What?) is designed to track people using a combination of learned textural appearance models and contour tracking (Haritaoglu et al., 2004). 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 whether 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 in their object model (Stauffer & Grimson, 2000).

2.3 *Behaviour analysis*

Object detection, tracking and classification, though unsolved problems in themselves, can be seen as precursors to the defining task of automated surveillance: the characterising of activity taking place in the scene, often called behaviour analysis. We now examine the most common applications of behaviour analysis, relating to human and traffic motion.

2.3.1 *Human motion analysis*

One of the most popular and demanding types of behaviour to analyse automatically is that of a human being (Aggarwal & Cai, 1999). 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 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, that of Oliver et al., 2000, 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). The system of Ivanov & Bobick, 2000, recognises simple human gestures (against a black backdrop), while an earlier system (Siskind & Morris, 1996) recognises simple actions (pick up, put down, push, pull, etc.), also based on a trained HMM. The detection of anomalous behaviour is addressed in Nair & Clark, 2002, 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 & Matsuyama, 2000, 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) that 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 (Ayers & Shah, 2001). 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 it requires no training.

A challenging problem under consideration in the UK is the detection of individuals with suicidal intent on London underground platforms. There is evident motivation to avoid the human tragedy, and additionally the enormous financial cost associated with temporary rupture of a transport artery. Whether such detection is possible, however, is unclear.

2.3.2 *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 & Vettnaker, 2000, 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 aforementioned Ivanov & Bobick system 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 is designed more specifically for the tracking and analysis of traffic from a static or vehicle mounted camera

(Ferryman et al., 2000). 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.

2.4 *Query-based video retrieval*

An application receiving growing attention is query-based video retrieval. Here the aim is to permit a user (e.g. security person) to ask questions of the system that require an automated search through recorded footage. Example queries might be:

- Which vehicle last parked in this bay?
- When did this person enter the airport?
- Who last entered this area?
- Who left this object here?
- Who did this person last meet?

Such a system evidently requires a means by which a user may select one of a legal set of query options, and query-specific information (e.g. an image subregion or a pointer to a person in a frame). The system should then trawl back through the recorded video in order to provide an appropriate answer.

Processing a video backwards in time has much in common with the more usual forwards processing. Indeed, there is no reason why query systems cannot be used in relation to future events, and in a sense standard surveillance systems do just that (except that the vocabulary is usually limited).

2.5 *Pan-Tilt-Zoom cameras*

So far the description of previous work has largely assumed that cameras have fixed characteristics. Additional opportunities arise when using a Pan-Tilt-Zoom (PTZ) camera for surveillance. Thus, for example, rather than just tracking an individual through a fixed FOV, it becomes possible to zoom in or out (so as to maintain, perhaps, an optimal-sized view of the face), and to rotate the camera to extend the effective FOV. Relatively little work has been done in the context of video surveillance, PTZ cameras having been studied most in the context of active vision in robotics.

3 A SURVEILLANCE SUCCESS STORY

In 1993 one of this paper's authors (Brooks) was asked by a high-technology startup company to develop a solution to the surveillance problem of automatically detecting unattended packages (e.g. suitcases in airports). This was taken up as a project within the CRC for Sensor Signal and Information Processing. A pilot implementation was generated using new technology based on keeping statistical properties of small windows of the video frame. The statistics enable characterization of the scene's background, even though it may never by visible in its entirety, possibly due to people constantly traversing the scene. The background image statistics are then recomputed and compared with the previous instance. A change in the background image statistics is associated with the entry of a new object into the scene that becomes stationary. For this reason, the system was termed a Background Change Detection method.

A broad overview of the system was published without revealing commercially sensitive details (Gibbins et al., 1996) and a patent was established (Brooks et al., 2001). After failing to gain commercial take-up for several years, the start-up company iOmniscient Pty Ltd licensed the technology in 2001 and commissioned further development by the inventors. This large-scale effort was concerned with the drive for reduced false positives and the adding of new features. Thus for example, a facility for specifying depth-sensitive detection-size thresholds was introduced.

In 2003, the resulting iOmniscient system won a federal and state governments award "Secrets of IT Innovation." In May 2004, the system won the prestigious 2004 International Fire & Security Exhibition and Conference award for the "Best Product in Intelligent Surveillance." Further international awards have since been won in Denmark, Taiwan and elsewhere.

Figure 1. Detection trial of suspicious objects in Federal Parliament House

Sales subsequently flowed, one of the first being to the Royal Ontario Art Gallery in Canada. There the system is used to protect paintings and other valuable artifacts in an efficient and cost effective manner, even though occlusions may be frequent. The system is now in use in airports and other major facilities and was recently chosen as a component of a comprehensive security facility for protecting the Sydney Harbour Bridge[2]. It has recently been on-

[2] See Australian Financial Review, 13 May 2005, p.3.

licensed to the company 2020 Vision Systems in the UK, amongst others.

The system owes its success in part to its possession of a valuable capability with a low rate of false positives and false negatives. Fig. 1 shows a frame from research footage in which two detections are highlighted. The system can be used in such diverse applications as detection of blocked road tunnels, computer theft from offices, graffiti, illegal parking, warehouse items that have not sold within a specified period, etc.

4 FULLY NETWORKED SURVEILLANCE

A common characteristic of most surveillance systems is that, although they may be implemented over a network of cameras, automated analysis is applied independently to each video stream. Thus, for example, the iOmniscient system described earlier analyses each video output in isolation. As it happens, this suffices for the applications in mind.

There is a strong need, however, for surveillance systems that are truly networked in that inferences can be made on the basis of the cooperative, real-time processing of multiple video streams generated simultaneously across the network.

A complexity of this problem is that a person walking through such a network may be viewed at any given time by a single camera, two or more cameras, or no camera. Fig. 2 shows a (hypothetical) plan view of part of an airport-terminal complex with single entry and exit ports. A network of 9 cameras is used to monitor the area. Shaded regions each depict a field of view (FOV) of a given camera. A person entering the terminal will obviously appear in camera 1 before camera 2. The unshaded area between the cameras indicates that the person will not be visible during the transition from one camera to the next. Cameras 3 and 4 have overlapping FOVs so that a person can be simultaneously visible in both cameras. A further complexity is that a person may be simultaneously visible in cameras 5, 6 and 7.

In Section 2.2, tracking was described in the context of a single camera (within-camera tracking). It was noted that progress has been rapid, but fully robust trackers have yet to be devised. Networked tracking refers to the problem of tracking an individual through a large network of cameras. Here the challenges are somewhat different and have received relatively little attention. A specific requirement might be as follows. A security person alerts the computer system to a given individual entering an airport and appearing in a particular camera's FOV. The system should then maintain a permanent display of this person, insofar as is possible. This will require the system to highlight the individual in some way as he moves around a given FOV. However, on leaving a FOV, the system will need to resume tracking as soon as possible in a neighbouring FOV. Complicating this problem is the fact that the person may at times be outside of any of the network FOVs.

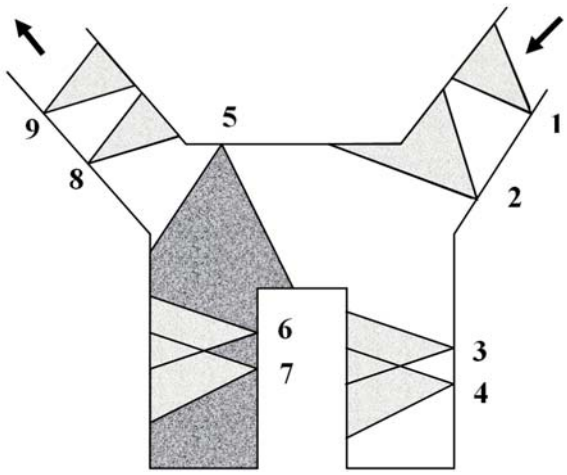

Figure 2. Camera fields of view within an airport complex

4.1 *Tracking using appearance information*

This camera handover problem is more complex than might be at first be realised. One reason for this is that using an entity's appearance alone is insufficient to correlate FOV transfers. Even in favourable circumstances cameras have differing characteristics that can result in appearance measures of the same individual being quite dissimilar. Compounding this problem is that part of the entity may be occluded in one FOV and not in the other, that illumination conditions might vary significantly between views, or that different views of the entity might be presented (e.g. front and back clothing appearances might be quite different). It is important to appreciate that appearance modeling at this stage does not involve high-resolution face detection methods or the like. Rather we are essentially operating with moving blobs and their colour and brightness characteristics. If tracking across the network is to be successful, additional sources of information will be required.

4.2 *Tracking using graph-based methods*

Since appearance information alone is not reliable enough to facilitate robust network tracking. An obvious strategy is to utilize a representation of the camera network itself, capturing:

- Camera FOV overlap
- The relationship between FOVs and neighbouring unviewed regions
- Statistics of time taken to transition from FOV to FOV, and FOV to unviewed region.
- Relative likelihood of certain transition options taking place.

The above information is naturally associated with a graph-based representation. As an initial candidate representation, consider a graph with each node containing one of the labels 3L, 5R, etc. Here the interpretation is as follows. Reaching a node 3L will indicate that the left part of the FOV of camera 3 has either been entered or exited. Figure 3 shows the entry-exit (EE) graph for the airport. Note that this captures the fact that a person may pass through the FOVs of cameras 3 and 4 in two ways—either sequentially, or through the overlapping FOV region. Also, the EE graph makes it clear that a 3R to 4R to 4L transition is impossible.

While this graph contains a significant amount of information, it needs careful interpretation, with a requirement that history information is maintained in order to infer the present state. For example, if during tracking we are situated in the graph at the 6R node, we are unable to determine whether this corresponds to our object actually being visible by camera 6 without recourse to previous states. Another problem is that multiple instances of the same node may appear, which in general would appear to render impossible the learning of such graphs from observations.

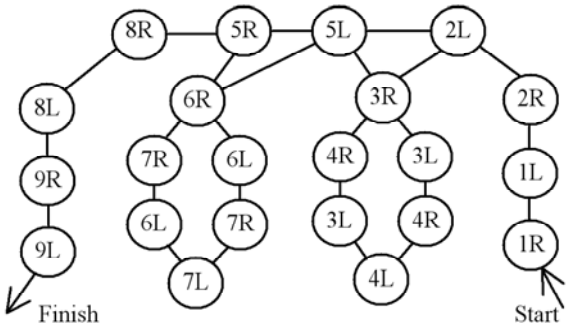

Figure 3. An entry-exit graph for the airport

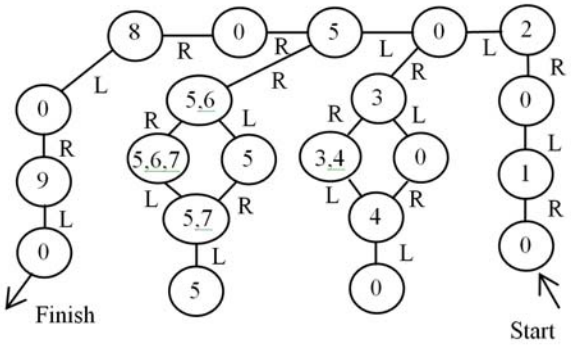

Figure 4. A visibility graph for the airport

An alternative approach is to employ a visibility graph shown in Figure 4. In this case a node holds a list of cameras within which the tracked object is presently visible. A value of zero indicates that the object is invisible. A transition to a neighbouring node occurs when an object either ceases to be visible or becomes visible by a camera. Links are labelled either L or R, indicating on which side of a FOV the transition takes place. This representation has the advantage of being much more explicit in its information content. Each node in this graph corresponds to an area of airport floor partitioned according to camera visibility. Note that statistics concerning the time taken to undergo a transition also need to be recorded on the links. Likewise, links exiting a given node can record the relative probability of the neighbouring transitions as determined from extensive observations previously undertaken. As transition probabilities are not necessarily symmetric, two relative probabilities will need to be recorded for each link.

Having acquired a graph-based structure capturing transition information, an effective tracking algorithm needs to be devised. One approach will be to incorporate full information into a Bayesian tracking formulation. Details will be left to a subsequent article.

4.3 *Networked behaviour analysis*

In a single camera FOV, behaviour analysis might involve detection of erratic or violent motion exhibited in a single video. Behaviour analysis across a network can involve detection of a range of other activities. Using network-tracking information, it becomes possible to determine whether an individual is moving through the domain in a standard manner. For example, in an airport, it may be important to isolate individuals that do not visit the check-in prior to passing through security barriers. Haphazard or backtracking paths might be highlighted. Persons repeatedly trying to open locked doors or looking into office windows might be of interest in an office complex. In a shopping centre, however, this behaviour may be of little interest, but groups of people repeatedly converging and splitting up might be tagged suspicious. Identifying these types of behaviours necessarily requires a networked approach to behaviour analysis.

4.4 *Query-based video retrieval*

As is the case for behaviour analysis there are certain video retrieval queries that are possible only in a networked video analysis system. After it has been determined that something has been stolen, for instance, it may be possible to identify the person responsible, after the fact, within the video sequence generated by a single camera. What networking adds to the system is the ability to see where else that person has been, what else they may have stolen, and where they are now.

4.5 *Pan-Tilt-Zoom in the network*

An interesting problem that has received some attention is the cooperative use of a multiple fixed cameras and a PTZ camera. Here, a large area may be monitored via several wide-angle low-cost cameras. In the event that they detect an object of interest, the PTZ camera can be used to obtain a relatively high quality, enlarged track of the object in question.

4.6 *Video surveillance platform*

The Video Surveillance and Analysis Group within the University of Adelaide has developed a networked surveillance platform which is currently monitoring internet cameras distributed across all areas of the campus. The system provides an interface to security personnel both within the security office and at hand-held units which may be carried into the field. The system also analyses the footage as it is generated by the cameras and archives segments which are identified as being of potential interest at a future date. The advantages of having this system installed and monitoring a real camera network for the group's research efforts are manifold. Current research into networked behaviour analysis and tracking is not only evaluated on real footage, but can also be tested by real users of the system.

5 FUTURE CHALLENGES

We conclude with some remarks concerning upcoming challenges in the areas of graph-based representations, scripting of videos and ubiquitous monitoring.

5.1 *Graph-based representations*

Rather than manually and tediously entering a graph representation for a specific surveillance application, it would be highly desirable to automatically acquire a representation of the camera network using connectivity data taken from individuals tracked through the network. This is something of a chicken and egg problem: we need a graph to aid tracking, but we need to do some tracking to build a graph. One solution would be to build the graph on the basis of tracking information gained (if and) when very few people traverse the environment, as appearance tracking can be effective. Even if we are able to do some accurate initial tracking, though, problems remain. If someone walks through the airport from the start to node 4 via 3,4 and back to 3 via the unobserved region, how are we to correctly link the instances of visibility in 3 rather than forming a graph that is simply a linear list of visited nodes? One strategy may be to utilise some natural structural properties of the representation, as arises, for example, with the special form for neighbouring cameras that overlap (see the regular structures in the above graphs). Nevertheless, this remains a significant hurdle.

Finally, how might we develop representations for more general situations in which objects may exit an FOV in a variety of ways (including, perhaps through a central area corresponding to a door)? The answer might lie in partitioning each FOV not into left and right regions but into regions based on visibility in other cameras. For each camera, this can be done by projecting the boundary of each overlapping FOV onto a common ground plane and then into the image of that camera (shown for camera 1 in Figure 5). This results in a set of lines denoting positions in the camera's image where an object moving in the ground plane appears or disappears from another camera's view. The position

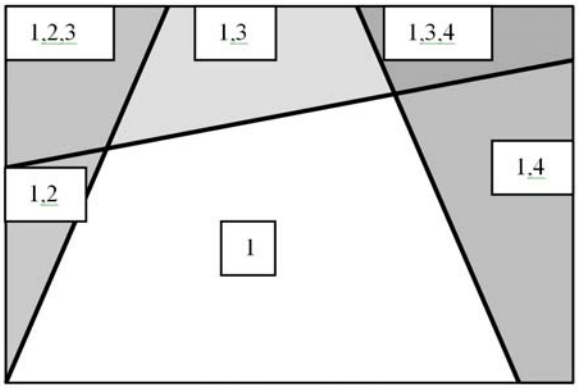

Figure 5. Camera 1 (unrelated to previous Figs) FOV divided according to co-visibility in 2, 3, 4

of an object in the image relative to these lines determines which other cameras can see the object, as shown in Khan et al., 2003. Each time an object crosses one of these boundaries, its state in the visibility graph changes. Khan et al. propose a method for determining these boundary lines without requiring camera calibration; instead it is achieved by observing the appearances and disappearances of objects from multiple FOVs. In a similar vein, Lee et al., 2000, describe another method for tracking across overlapping FOVs that does not require camera calibration, provided all cameras can see a common ground plane. The algorithm works by aligning observed trajectories from each camera, assuming each trajectory is within the ground plane. In this way the cameras' overlap and relative pose can be recovered. By acquiring these overlapping regions, these methods define the connectivity of the camera graph for overlapping cameras, which is an important step towards the automatic acquisition of the graph.

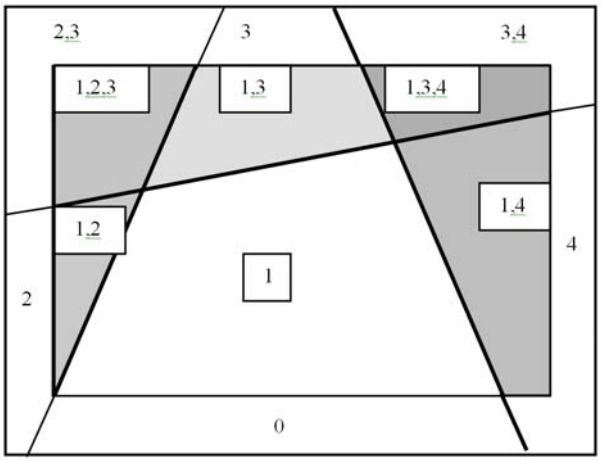

Figure 6. Camera 1 FOV, with peripheral regions

To extend these techniques to non-overlapping cameras, one possibility is to introduce a peripheral region surrounding each camera's FOV (Figure 6). This region is divided by the same FOV boundaries defined within the camera's FOV. If the object is still visible to at least one other camera when it is in this peripheral band, that region of the periphery is assigned the same label as the region in which it is seen in the other camera. If the object is not visible to any camera, it is assigned a unique label (0 in Figure 6). In most cases, this region will also border another camera. In the airport example, the invisible region to the left of camera 2 is the same as the region to the right of camera 3. An interesting problem will be to match these regions across cameras.

5.2 *Scripting of video for fast retrieval*

A concept at present in its infancy is the real-time scripting of incoming video. The intention here is to describe the activity in a video at a relatively high level, possibly in textual form. This might involve descriptions such as "red object slowly enters FOV from left and comes to halt at FOV centre", or "two entities meet at FOV top right before exiting FOV together". The aim of this scripting is to facilitate rapid search of the video as might be required in a query-based event detection system. Clearly, the textual description itself can then be searched, which in turn will link into the video. This has been termed the "Inverse Hollywood" problem in that it goes from video to script, rather than script to video.

Scripting opens up fascinating possibilities. For example, suppose that several thousand cameras monitor a city's streets and that the associated videos are stored in digital repositories that are networked together. If a bank robber's car is now viewed in one of the cameras, how might the escape path through the city be traced? The challenge is then to search at two levels: a meta-search is conducted whereby candidate repositories are selected, and a script search is undertaken with respect to an individual video.

5.3 *Ubiquitous monitoring*

The number of cameras required to completely monitor all but the smallest of spaces is surprisingly large. This is due to the obvious fact that cameras cannot see around corners, but also because the sensor resolution is relatively low. This means that people only tens of metres from the camera occupy so few pixels that any form of identification becomes very difficult. Increasing the number of cameras, however, requires processing that many more video streams. This expansion can obviously only take place up to a limit. New camera technologies are being developed which will dramatically increase the number of cameras that may be installed in a given area. What has yet to be determined is how to process effectively the volume of video generated.

REFERENCES

Aggarwal, J.K. & Cai. Q. 1999. Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428-440, March.

Ayers, D. & Shah, M. 2001. Monitoring human behavior from video taken in an office environment. *Image and Vision Computing*, 19(12):833–846, October.

Beardsley, P.A., Zisserman, A. & Murray, D.W. 1997. Sequential updating of projective and affine structure from motion. *International Journal of Computer Vision*, 23(3):235–259.

Brand, M. & Kettnaker, V. 2000. Discovery and segmentation of activities in video. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):844–851, August.

Brooks, M.J., Gibbins, D. & Newsam, G. 2001. Non-Motion Detection, PCT/AU02/01574, WO2003/ 044752, Priority Date: 21 Nov. 2001; Filing Date: 21 Nov. 2002.

Cutler, R. & Davis, L.S. 2000. Robust real-time periodic motion detection, analysis, and applications. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):781–796, August.

Ferryman, J.M. Maybank, S.J. & Worrall, A.D. 2000. Visual surveillance for moving vehicles. *International Journal of Computer Vision*, 37(2):187–197, June.

Gibbins, D., Newsam, G. & Brooks, M.J. 1996. Detecting suspicious background changes in video surveillance of busy scenes. In *Third IEEE Workshop on Applications of Computer Vision*, pages 22–26.

Haritaoglu, I., Harwood, D. & Davis, L.S. 2000. W4: Real-time surveillance of people and their activities. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):809–830, August.

Isard, M. & Blake, A. 1998. Condensation-conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5-28.

Ivanov, Y.A. & Bobick, A.F 2000. Recognition of visual activities and interactions by stochastic parsing. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):852–872, August.

Javed, O., Shafique, K. & Shah, M. 2002. A hierarchical approach to robust background subtraction using color and

gradient information. In *Workshop on Motion and Video Computing (MOTION'02)*, pages 22–27.

Khan, S. & Shah, M. 2003. Consistent labeling of tracked objects in multiple cameras with overlapping fields of view. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(10):1355-1360, October.

McIvor, A., Zang, Q. & Klette, R. 2000. The background subtraction problem for video surveillance systems. Technical Report CITR-TR-78, University of Auckland.

Nair, V. & Clark, J.J. 2002. Automated visual surveillance using hidden markov models. In *International Conference on Vision Interface*, pages 88–93.

Oliver, N.M., Rosario, B. & Pentland, A.P. 2000. A Bayesian computer vision system for modeling human interactions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):831–843, August.

Pope, A.R. 1994. Model-based object recognition: A survey of recent research. Technical Report 94-04, University of British Columbia.

Raja, Y., McKenna, S.J. & Gong, S. 1998. Tracking and segmenting people in varying lighting conditions using colour. In *IEEE International Conference on Face and Gesture Recognition*, pages 228–233.

Siskind, J.M. & Morris, Q. 1996. A maximum-likelihood approach to visual event classification. In *Proc. 4th European Conference on Computer Vision*, pages II:347–360.

Stauffer, C. & Grimson, W.E.L. 2000. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757.

Velastin, S., Sanchez-Svensson, M., Sun, J., Vicencio-Silva, M., Aubert, D., Lemmer, A., Brice, P., Khoudor, L. & Kallweit, S. 2003. D7P: Innovative tools for security in transports. Technical Report GRD1-2000-10601, PRISMATICA Project, 5th Framework Programme.

Wada, T. & Matsuyama, T. 2000. Multiobject behavior recognition by event driven selective attention method. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):873–867, August.