Automated Visual Surveillance in computer vision

B ANANDAMPILAI
Dept. of computer engg. Noorul islam college of engineering, kumarakooil, nagerkoil

K.KRISHNAMOORTHY
ASST.PROF/COMPUTER SCIENCE & ENGG
MVJ COLLEGE OF ENGINEERING, BANGAORE
Email- drbapilai@gmail.com

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 analyze behavior automatically. This paper reviews progress in the field and considers some of the major remaining problems in automated video surveillance.

Keywords: Automated visual surveillance, CCTV, VSAM, Behavior analysis, Traffic motion analysis, Surveillance video mining.

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. 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 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 behavior requires a much richer description of an individual, possibly including a history of their overall motion, limb movements and gaze direction. Although the exact requirements vary between surveillance systems, there are issues that are common to all. 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. Once these objects of interest have been identified, a subsequent problem is that of behavior analysis: describing their activity in such a way that a course of action can be decided upon in software. Object identification and tracking, and behavior 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. 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. 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. Alternatively “snake” contours can be used to detect an object outline and then track.

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. Pixel or region based background subtraction techniques, in which a model of foreground and/or background appearance is learnt; blob
trackers have also been used. A Bayesian approach to region based object detection and tracking without background subtraction is presented in .

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 parameterized on width and height is often used. For example, W4 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 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 ). 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. 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, 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, 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), Quadra pedal (dogs) and “other” objects from some aerial footage.

3 Behavior 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 behavior analysis.

3.1 Human motion analysis

One of the most popular and demanding types of behavior to analyze automatically is that of a human being. 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 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 recognizes simple human gestures (against a black backdrop), while an earlier system recognizes simple actions (pick up, put down, push, pull, etc.), also based on a trained HMM. The detection of anomalous behavior is addressed in 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. who use a “hypothesis and test” algorithm to interpret and predict human behavior 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 recognize a limited set of human behaviors. 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 modeled before any data is seen, but requires no training.

Rather than using a HMM to learn and then recognize human activity, Yacoob and Black 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 recognize. 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 recognized in subsequent footage by minimizing a simple sum of squared distance measure. This algorithm is demonstrated
recognizing 4 different types of walk, using about 15 features, with about an 80% success rate. Mann et al. Analyze video content by physically modeling 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 the observed behavior is reported in natural language sentences rather than as a set of disjointed phrases. Behavior is recognized, 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 behavior recognition systems, however, the range of recognized 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 behavior and motion of traffic in an urban area is quite different in nature to human behavior, similar techniques apply to its analysis. Brand and Vettnaker [4] present a system that learns patterns of behavior by training a HMM. Anomalous behavior, 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 behavior (such as falling asleep or standing at the window, apparently). The 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. 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.

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. 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. Indoors 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 omniscient commercial system 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 require both camera calibration and overlapping fields of view to maintain tracks across cameras. Additionally, a 3D model of the environment is required by 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 information, but still require overlapping views to establish correspondences. Kettnaker 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 that it is impractical to monitor entirely. A system using cameras situated about 2 miles apart along a highway is described in . 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 minimize 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.

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. 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 databases, 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 place, it is common for police to scan (by eye) numerous videotapes 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. For example, a system may be required to search a video repository’s last few days’ footage for instances of violent behavior, or examples of where a car passed unusually quickly through the field of view.

If several thousand networked digital video repositories were 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, and remains a tremendous challenge.

6 Conclusion
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 realized. 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 unfavorable weather and lighting conditions. However progress is being made ever more rapidly, and the demand for automated surveillance continue 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”.

References


