ICICS-PCM 2003
15-18 December 2003
Singapore

Smart Surveillance: Applications, Technologies and Implications

Arun Hampapur, Lisa Brown, Jonathan Connell, Sharat Pankanti, Andrew Senior and Yingli Tian.

IBM T.J. Watson Research Center,
19 Skyline Drive, Hawthorne, NY 10532
arunh@us.ibm.com

Abstract

Smart surveillance, is the use of automatic video analysis technologies in video surveillance applications. This paper attempts to answer a number of questions about smart surveillance: What are the applications of smart surveillance? What are the system architectures for smart surveillance? What are the key technologies? What are the some of the key technical challenges? and What are the implications of smart surveillance, both to security and privacy?

1. Introduction

Recent world events have created a shift in the security paradigm from "*investigation of incidents*" to "*prevention of potentially catastrophic incidents*". Existing digital video surveillance systems provide the infrastructure only to capture, store and distribute video, while leaving the task of threat detection exclusively to human operators. Human monitoring of surveillance video is a very labor-intensive task. It is generally agreed that watching video feeds requires a higher level of visual attention than most every day tasks. Specifically vigilance, the ability to hold attention and to react to rarely occurring events, is extremely demanding and prone to error due to lapses in attention [12]. One of the conclusions of a recent study by the US National Institute of Justice [6], into the effectiveness of human monitoring of surveillance video, is as follows

*"These studies demonstrated that such a task[..manually detecting events in surveillance video], even when assigned to a person who is dedicated and well-intentioned, **will not support an effective security system**. After only 20 minutes of watching and evaluating monitor screens, the attention of most individuals has degenerated to well below acceptable levels. Monitoring video screens **is both boring and mesmerizing**. There are no intellectually engaging stimuli, such as when watching a television program."*

Clearly today's video surveillance systems while providing the basic functionality fall short of providing the level of information need to change the security paradigm from "investigation to preemption". Automatic visual analysis technologies can move today's video surveillance systems from the investigative to preventive paradigm. Smart Surveillance Systems provide a number of advantages over traditional video surveillance systems, including

- "the ability to preempt incidents -- through real time alarms for suspicious behaviors"
- "enhanced forensic capabilities -- through content based video retrieval"
- "situational awareness – through joint awareness of location, identity and activity of objects in the monitored space".

Section 2 provides a short introduction to various applications of smart surveillance systems. Section 3 discusses the architectures for smart surveillance systems. Section 4 presents the key technologies for smart surveillance systems. Section 5 briefly discusses the challenges in smart surveillance. Section 6 discusses the implications of smart surveillance technologies. Conclusions are presented in section 7.

2. Applications of Smart Surveillance

In this section we describe a few applications of smart surveillance technology. In this section, we describe a few applications. We group the applications into three broad categories, real time alerts, automatic forensic video retrieval, and situation awareness.

2.1 Real Time Alerts: There are two types of alerts that can be generated by a smart surveillance system, user defined alerts and automatic unusual activity alerts.

2.1.1 User Defined Alerts: Here the system is required to recognize a variety of user defined events that occur in the monitored space and notify the user in real time, thus providing the user with an opportunity to evaluate the situation and take preventive action if necessary. Following are some typical events.

1. Generic Alerts: These are alerts which depend solely on the movement properties of objects within the monitored space. Following are a few common examples.

1. Motion Detection: This alert detects movement of any object within a specified zone.
2. Motion Characteristic Detection: These alerts detect a variety of motion properties of objects, including specific direction of object movement (entry through exit lane), object velocity bounds checking (object moving too fast).
3. Abandoned Object Alert: This detects objects which are abandoned, e.g., a piece of unattended baggage in an airport, or a car parked in a loading zone.
4. Object Removal: This detects movements of a user-specified object that is not expected to move, for example, a painting in a museum.

2. Class Specific Alerts: These are alerts which use the type of object in addition to the object's movement properties. Following are a few common examples.

1. Type Specific Movement Detection: Consider a camera that is monitoring runways at an airport. In such a scene, the system could provide an alert

0-7803-8185-8/03/$17.00 © 2003 IEEE

on the presence or movement of people on the tarmac but not those of aircrafts.

2. Statistics: Example applications include, alerts based on people counts (e.g., more than one person in security locker) or people densities (e.g., discotheque crowded beyond an acceptable level).

3. Behavioral Alerts: These alerts are generated based on adherence to, or deviation from, learnt models of motion patterns. Such models are typically trained by analyzing movement patterns over extended periods of time. These alerts tend to be very application specific and use a significant amount of context information, for example,

1. Detecting shopping groups at retail checkout counters, and alerting the store manager when the length of the queue at a counter exceeds a specified number. [8]
2. Detecting suspicious behavior in parking lots, for example, a person stopping and trying to open multiple cars.

4. High Value Video Capture: This is an application which augments real time alerts by capturing selected clips of video based on pre-specified criteria. This becomes highly relevant in the context of smart camera networks which use wireless communication.

2.1.2 Automatic Unusual Activity Alerts: Unlike the user defined alerts, here the system generates alerts when it detects "activity that deviates from the norm". The smart surveillance system achieves this based on "learning" normal activity patterns [17]. For example, a smart surveillance system that is monitoring a street learns that "vehicles move about on the road" and "people move about on the side walk". Based on this pattern the system will provide an alert when a car drives on the side walk. Such unusual activity detection is the key to effective smart surveillance, as all the events of interest cannot be manually specified by the user.

2.2 Automatic Forensic Video Retrieval (AFVR):

The capability to support forensic video retrieval is based on the rich video index generated by automatic tracking technology. This is a critical value-add from using smart surveillance technologies. Typically the index consists of such measurements as object shape, size and appearance information, temporal trajectories of objects over time, object type information, in some cases specific object identification information. In advanced systems, the index may contain object activity information. The Washington, DC sniper incident is a prime example of where AFVR could be a break-through technology. During the incident the investigative agencies had access to hundreds of hours of video surveillance footage drawn from a wide variety of surveillance cameras covering the areas in the vicinity of the various incidents. However, the task of manually sifting through hundreds of hours of video for investigative purposes is almost impossible. However if the collection of videos were indexed using visual analysis, it would enable the following ways of retrieving the video

1. Spatio-Temporal Video Retrieval: An example query in this class would be, "Retrieve all clips of video where a "blue car" drove in front of the "7/11 Store on 23rd street" between the "26th of July 2pm and 27th of July 9am" at "speeds > 25mph".

2. Surveillance Video Mining: In the case of the Washington sniper incident, the surveillance video mining application would attempt to present the users with a set of potential movement patterns of cars over a set of cameras covering multiple incident locations, this would enable the investigative agencies to answer questions like "Was there a single car that appeared in all of the incident locations?".

2.3 Situation Awareness:

Ensuring total security at a facility requires systems that can perpetually track the identity, location and activity of people and vehicles within the monitored space. For example, the existing surveillance technology cannot answer questions such as: did a single person loiter around near a high security building on multiple occasions? Such perpetual tracking can be the basis for very high levels of security. Typically surveillance systems have focused on tracking location and activity, while biometrics systems have focused on identifying individuals. As smart surveillance technologies mature [7], it becomes possible to address all these three key challenges in a single unified frame work giving rise to, *joint location identity and activity awareness*, which when combined with the application context becomes the basis for situation awareness.

3. Smart Surveillance Architectures:

In this section we discuss how smart surveillance technologies are incorporated into a complete surveillance system. We discuss three different types of smart surveillance architectures

3.1 Basic Smart Surveillance Architecture (BSSA):

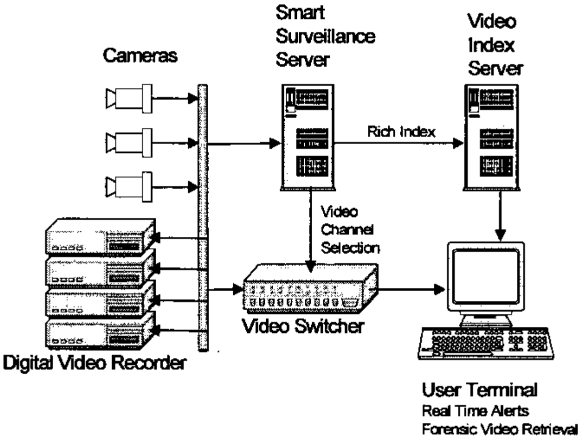

Figure 1: Block diagram of a smart surveillance system.

Figure 1 shows the block diagram of a basic smart surveillance system. The outputs of video cameras are recorded digitally and simultaneously analyzed by the smart surveillance server, which produces real time alerts and a rich video index. The types and parameters of the alerts are user configurable. The user can use the rich index to retrieve video from the archive for forensic. In the BSSA the video recording and analysis is centralized requiring the cameras to be wired to a central location. The role of automatic visual analysis in the BSSA is primarily analytical.

3.2 Active Smart Surveillance Architecture (ASSA):

Figure 2 shows the block diagram for an active smart surveillance architecture. The key difference between this and a BSSA is the addition of active camera controls. Here the automatic visual analysis is used not only to "understand what is going on in the scene" but also "to selectively pay more attention" to automatically detected activities or events of interest. The ASSA could be used for many different applications. Below we describe two examples.

1. Face Catalogler: This is a system which aims to **non intrusively** acquire a high-resolution face images of all people passing through a space, Here ASSA detects and tracks people and uses the active cameras to zoom in and acquire high resolution face pictures.
2. Multi-scale Video: This is a system which automatically allocates higher resolution to portions of the scene which have certain predetermined types of activity. For example, all cars that are moving with high speed through a parking lot may be imaged at a higher resolution through the active cameras.

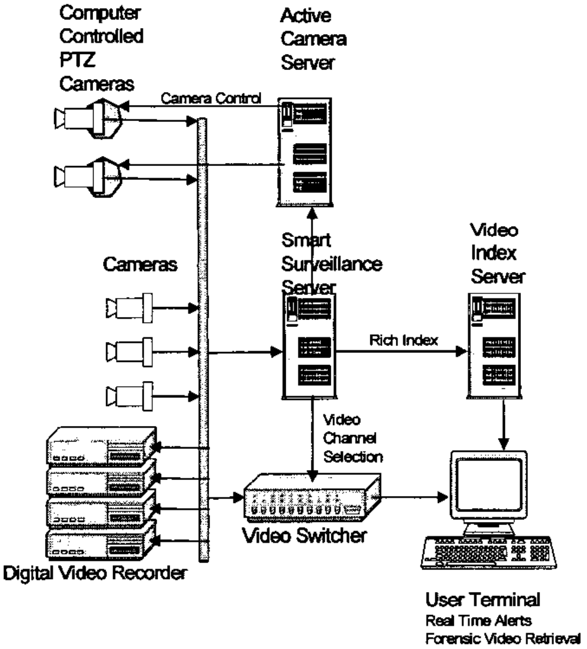

Figure 2: Block diagram of an Active Smart Surveillance System.

The key technical difference between the ASSSA and BSSA are

- Active Camera Resource Management: This includes techniques for deciding which of the active cameras are to be used to meet the overall goals of the system.
- Active Camera Image Analysis: This includes analysis of the active camera images in order to control the movement and zoom of the cameras.

3.3 Distributed Smart Surveillance Architecture (DSSA)

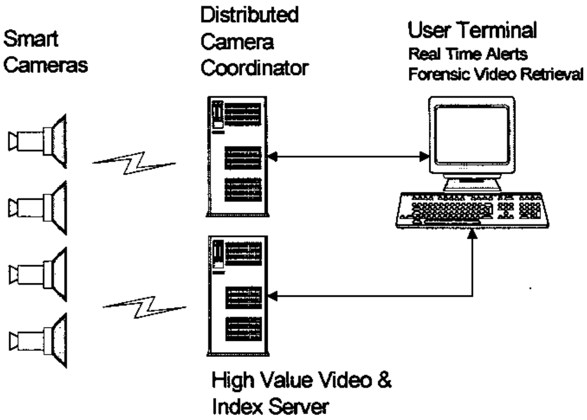

Figure 3: Block diagram of a distributed smart surveillance architecture using smart cameras.

One of the biggest costs in deploying a surveillance system is the infrastructure (e.g., wiring) required to move the video from the cameras to a central location where it can be analyzed and stored. Further, the wiring capacity is not easily scalable and because of involvement of manual labor, instantaneous/portable installation of network is difficult. There is an increasing trend to building cameras which in addition to generating images also analyze them with on-board processing. Such cameras are typically called **smart camera.** The goal of a smart camera system is to minimize the cost of deployment. In such architecture (fig 3), the camera would typically use wireless communication to coordinate with a central camera coordinator. A smart camera would use the automatic visual analysis to determine how to use the limited storage and bandwidth to effectively transmit only "high value video" to the server.

4. Technologies for Smart Surveillance.

Among the three architectures presented in the previous section, the BSSA is likely to be the most ubiquitous in the near future. The BSSA is an enhancement of the architecture of current day surveillance systems, which have cameras mounted in different parts of a facility, all of which are wired to a central control room. In this section, we will discuss various technologies that enable the BSSA and discuss the challenges involved as such systems get widely deployed.

Figure 4 shows the internal structure of a smart surveillance server which is one of the key components in the BSSA. The video from a camera is processed to detect moving objects of interest, these objects are tracked as they move about in the monitored space. The tracked object becomes a fundamental internal representation in the system upon which a number of processes act, including classifications and real time alert module. In this section we discuss each of the key components of the smart surveillance server.

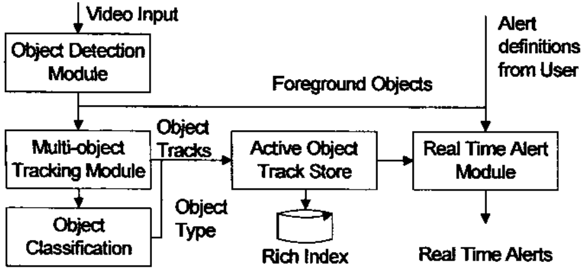

Figure 4: Internal structure of the smart surveillance engine.

4.1 Object Detection:

Most of the work on object detection relies heavily on the assumption of a static camera [9]. There is some work which has looked at detecting independent motion in moving camera images where the camera motion is well modeled [18]. Our approach to object detection has been two pronged each of which we discuss briefly below.

4.1.1: Adaptive Background Subtraction with Healing: The background subtraction module combines evidence from differences in color, texture, and motion. The use of multiple modalities improves the detection of objects in cluttered environments. The resulting saliency map is smoothed using morphological operators and then small holes and blobs are eliminated to generate a clean foreground mask. The background subtraction module has a number of mechanisms to handle changing ambient conditions and scene composition. First, it continually updates its overall RGB channel noise parameters to compensate for changing light levels. Second, it estimates and corrects for AGC (automatic gain control) and AWB (automatic while balance) shifts induced by the camera. Thirdly, it maintains a map of high activity regions and slowly updates its background model only in areas deemed as relatively quiescent. Finally, it automatically eliminates occasional spurious foreground objects based on their motion patterns.

4.1.2: Salient Motion Detection: This is a complementary approach to background subtraction. Here we approach the problem from a motion filtering perspective. Consider the following figure 5, the image on the left shows a scene where a person is walking in from of a bush which is waving in the wind. The next image in figure 5 shows the output of a traditional background subtraction algorithm (which per its design correctly classifies the entire bush as a moving object). However, in this situation, we are interested in detecting the person as opposed to the moving bush. Our approach uses optical flow as the basis for detecting salient motion. We use a temporal window of N frames (typically 10-15) to assess the coherence of optic flow at each pixel over the entire temporal window. Pixels with coherent optical flow are labeled as candidates. The candidates from the motion filtering are then subjected to a region growing process to obtain the final detection.

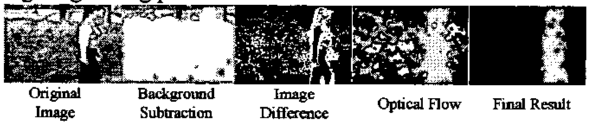

Figure 5 Illustration of salient motion detection.

Background subtraction and salient motion detection are complementary approaches, each with its strengths and weakness. Background subtraction is more suited for indoor environments where lighting is fairly stable and distracting motions are limited, whereas salient motion detection is well suited to detect coherent motion in outdoor situations.

4.2 Multi-Object Tracking:

Multi-object tracking attempts to associate objects with one another over time, by using a combination of the objects appearance and movement characteristics. This has been a very active area of research [2,3,4,5,8] in the past several years. Our approach to multi-object tracking has focused heavily on handling occlusions [15]. Following is a brief description of our approach.

The multi-object blob tracking relies on appearance models which are image-based templates of object appearance. New appearance models are created when an object enters a scene. In every new frame, each of the existing tracks is used to try to explain the foreground pixels. The fitting mechanism used is correlation, implemented as minimization of sum of absolute pixel differences between the detected foreground area and an existing appearance model. During occlusions, foreground pixels may represent appearance of overlapping objects. Color similarity is used to determine occlusion information (relative depth ordering) for the object tracks. Once this relative depth ordering is established, the tracks are correlated in order of depth. The correlation process is gated by the explanation map which holds at each pixel the identities of the tracks explaining the pixels. Thus foreground pixels that have already been explained by a track do not participate in the correlation process with models of the objects which are more distant. The explanation map is now used to update the appearance models of objects associated with each of the existing tracks. Regions of foreground pixels that are not explained by existing tracks are candidates for new tracks. A detailed discussion of the 2D multi-blob tracking algorithm can be found in [15]. *The 2D multi- object tracker is capable of tracking multiple objects moving within the field of view of the camera, while maintaining an accurate models of the shapes and colors of the objects.*

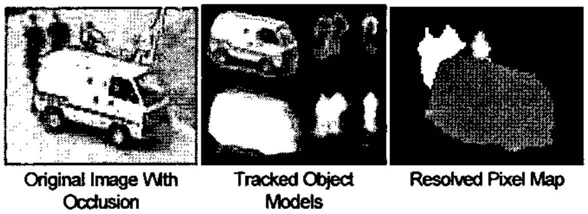

Figure 6: Occlusion handling in object tracking: Left Original image with occlusion. Middle: Objects being tracked by the system. Right: Classification of pixels into models.

4.3 Object Classification:

Moving foreground objects are classified into relevant categories. Statistics about the appearance, shape, and motion of moving objects can be used to quickly distinguish people, vehicles, carts, animals, doors opening/closing, trees moving in the breeze, etc. Our system classifies objects into vehicles, individuals, and groups of people based on shape features (compactness and ellipse parameters), recurrent motion measurements, speed and direction of motion (see Fig 7). From a small set of training examples, we are able to classify objects in similar footage using a Fisher linear discriminant classfier and temporal consistency information.

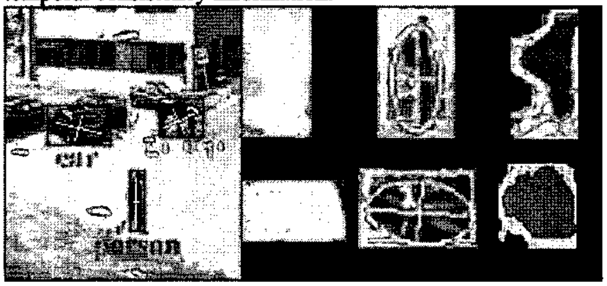

Figure 7: Left: Output of object classification algorithm. Right Intermediate steps in object classification

4.4 Real Time Alert Module:

The real time alert module uses the information produced by the other modules, namely, object detection, tracking and classification to detect user specified alerts. The key feature of the real time alert module is its extensible design. Figure 8 shows the generic structure of the real time alert module. This structure is instantiated by most of the alert types. In order to illustrate the structure of the module we present the design of the directional motion alert as an example. The following processes are instantiated when the user specifies a directional motion alert.

1. Directional Motion Alert Manager: Each motion alert user definition instantiates a directional motion alert manager, which is responsible for ensuring correct monitoring of the scene. The alert manager ensures that for every object being tracked there is a corresponding Object Track Observer that is instantiated. And that the Object Track Observer is deleted immediately upon the exit of the object from the scene.

2. Directional Motion Object Track Observer: This is the process that is charged with the job of measuring the direction of motion of the object and comparing it to the user specified direction. When ever the object motion direction matches the user specified direction, the object track observer issues a real time alert. The application uses the alert to signal the user that one of the specified alert conditions has been met.

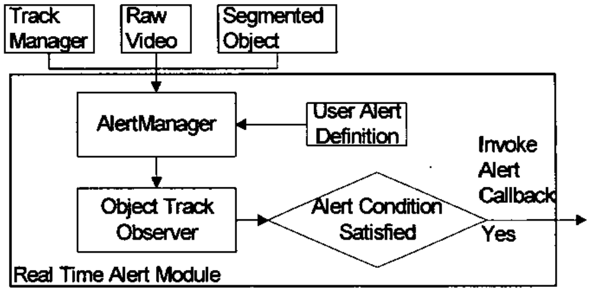

Figure 8: Internal structure of the real time alert module.

The exact nature of the object track observer depends on the particular alert it is implementing. However the general structure of many types of alerts is similar to that described above.

5. Challenges

There are two types of challenges that we highlight in the future development of smart surveillance systems.

1. **Technical Challenges:** There are a number of technical challenges that still need to be addressed in the underlying visual analysis technologies. These include challenges in robust object detection, tracking objects in crowded environments, challenges in tracking articulated bodies for activity understanding, combining biometric technologies like face recognition with surveillance to achieve situation awareness.

2. **Challenges in Performance Evaluation:** This is a very significant challenge in smart surveillance system. Evaluating performance of video analysis systems requires significant amounts of annotated data. Typically annotation is a very expensive and tedious process. Additionally, there can be significant errors in annotation. All of these issues make performance evaluation a significant challenge.

6. Implications of Smart Surveillance

Smart surveillance is a technology that has many different applications and potentially has significant implications to each of these. We look at implications primarily in the surveillance application, namely, security and privacy.

Security Implications: Clearly, the ability to provide real time alerts, capture high value video and provide sophisticated forensic video retrieval has the potential to enhance security in various public and private facilities. However, the value of the technology is yet to be proven in the field. As more and more smart surveillance systems get deployed the exact value will be known. In particular, systems must be analysed for their effectiveness in detecting events of interest, while generating few false alarms. In the first instance smart surveillance systems are intended to assist security guards, and will be measured on their ability to improve vigilance and to reduce labor and storage costs.

Privacy Implications: Smart surveillance systems have the ability to monitor video at a level which is humanly impossible. This provides the monitoring agencies with a significantly enhanced level of information about the people in the space leading to higher concerns about individual privacy and abuse of sensitive individual information. However, the same smart surveillance technology by virtue of indexing the video provides novel ways of enhancing privacy in video based systems which was hitherto not possible. Further details on the privacy preserving aspects of smart surveillance technologies can be found in [16].

7. Conclusions

While it is difficult to foresee a future where the surveillance of the monitored space is completely automatic, there is clearly an urgent need to augment the existing surveillance technology with better tools to aid efficacy of the human operators. With the increasing availability of the inexpensive computing, video infrastructure and better video analysis technologies smart surveillance systems will completely replace existing surveillance systems. The "degree of smartness" will vary with the level of security offered by such systems.

8. Acknowledgements

All the technical discussions in this paper are based on ongoing work in visual tracking at IBM research. The authors wish to acknowledge the PeopleVision project at IBM T.J.Watson Research Center [14].

References

1. Lisa Brown and Yingli Tian, Comparative Study of Coarse Head Pose Estimation, IEEE Workshop on Motion and Video Computing, Orlando FL, Dec. 5-6, 2002
2. Anjum Ali, J. K. Aggarwal: Segmentation and Recognition of Continuous Human Activity. IEEE Workshop on Detection and Recognition of Events in Video 2001.
3. Collins, Lipton, Fujiyoshi, and Kanade, "Algorithms for cooperative multisensor surveillance," Proc. IEEE , Vol. 89, No. 10, Oct. 2001.
4. D. Comaniciu, V. Ramesh, P. Meer: Real-Time Tracking of Non-Rigid Objects using Mean Shift,, IEEE Conf. Computer Vision and Pattern Recognition (CVPR'00), Hilton Head Island, South Carolina, Vol. 2, 142-149, 2000
5. Trevor Darrell, David Demirdjian, Neal Checka, Pedro Felzenszwalb: Plan-View Trajectory Estimation with Dense Stereo Background Models. ICCV 2001: 628-635.
6. Mary W. Green, The Appropriate and Effective Use of Security Technologies in U.S. Schools, A Guide for Schools and Law Enforcement Agencies, Sandia National Laboratories, September 1999, NCJ 178265
7. A. Hampapur et al., Face Cataloger: Multi-Scale Imaging for Relating Identity to Location, IEEE International Conference on Advanced Video and Signal Based Surveillance, Miami, FL, July 03.
8. Haritaoğlu and Flickner, "Detection and Tracking of Shopping Groups in Stores," CVPR 2001.
9. T. Horprasert, D. Harwood, and L. Davis. A Statistical Approach for Real-Time Robust Background Subtraction and Shadow Detection. Proceedings of IEEE Frame-Rate Workshop, Kerkyra, Greece, 1999.
10. A. K. Jain, R. Bolle, and S. Pankanti (eds.), Biometrics: Personal Identification in Networked Society, Kluwer Academic, 1999. (ISBN 0792383451)
11. G. Kogut, M. Trivedi, "A Wide Area Tracking System for Vision Sensor Networks," 9th World Congress on Intelligent Transport Systems, Chicalgo, Illinois,October,2002.
12. Miller et al, Crew fatigue and performance on US coast guard cutters, Oct 1998, US Dept of Transportation.
13. Anurag Mittal and Larry S. Davis, M2Tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene. International Journal of Computer Vision. Vol. 51 (3), Feb/March 2003.
14. PeopleVision: Demo videos of on going work at IBM research: www.research.ibm.com/peoplevision
15. A. Senior et al., Appearance Models for Occlusion Handling. Proceedings of IEEE Workshop on Performance Evaluation of Tracking and Surveillance Systems, 2001.
16. A Senior et al, Enabling video privacy through computer vision, To appear, IEEE Security and Privacy Magazine.
17. C. Stauffer, Automatic hierarchical classification using time-based co-occurrences, IEEE Conf on CVPR99, 333--339, 1999,
18. H. Tao, H.S. Sawhney and R. Kumar, Dynamic Layer Representation with Applications to Tracking, Proc. of the IEEE Computer Vision & Pattern Recognition, Hilton Head, SC, 2000.