Study of Robust and Intelligent Surveillance in Visible and Multi-modal Framework

Praveen Kumar, Ankush Mittal and Padam Kumar
Department of Electronics and Computer Engineering,
Indian Institute of Technology, Roorkee,
India 247667
E-mail: praveen.kverma@gmail.com,{ankumfec,padamfec}@iitr.ernet.in

Keywords: video surveillance, object detection and tracking, data fusion, event detection

Received: November 12, 2007

This paper gives a review of current state of the art in the development of robust and intelligent surveillance systems, going beyond traditional vision based framework to more advanced multi-modal framework. The goal of automated surveillance system is to assist the human operator in scene analysis and event classification by automatically detecting the objects and analyzing their behavior using computer vision, pattern recognition and signal processing techniques. This review addresses several advancements made in these fields while bringing out the fact that realizing a practical end to end surveillance system still remains a difficult task due to several challenges faced in a real world scenario. With the advancement in sensor and computing technology, it is now economically and technically feasible to adopt multi-camera and multi-modal framework to meet the need of efficient surveillance system in wide range of security applications like security guard for communities and important buildings, traffic surveillance in cities and military applications. Therefore our review includes significant discussion on multi-modal data fusion approach for robust operation. Finally we conclude with discussion on possible future research directions.

Povzetek: Podan je pregled metod inteligentnega video nadzora.

1 Introduction

Security of human lives and property has always been a major concern for civilization for several centuries. In modern civilization, the threats of theft, accidents, terrorists' attacks and riots are ever increasing. Due to the high amount of useful information that can be extracted from a video sequence, video surveillance has come up as an effective tool to forestall these security problems. The automated security market is growing at a constant and high rate that is expected to sustain for decades [1]. Video surveillance is one of the fastest growing sectors in the security market due to its wide range of potential applications, such as a intruder detection for shopping mall and important buildings [2], traffic surveillance in cities and detection of military targets[3], recognition of violent/dangerous behaviors (eg. in buildings, lifts) [4] etc. The projections of the compound annual growth rate of the video-surveillance market are about 23% over 2001-2011, to touch US$670.7 million and US$188.3 million in USA and Europe, respectively [5].

An automated surveillance system attempts to detect, recognize and track objects of interest from video obtained by cameras along with information from other sensors installed in the monitored area. The aim of an automated visual surveillance system is to obtain the description of what is happening in a monitored area and to automatically take appropriate action like alerting a human supervisor, based on the perceived description. Visual surveillance in dynamic scenes, especially for humans and vehicles, is currently one of the most active research topics in computer vision [6]. For at least two decades, the scientific community has been involved in experimenting with video surveillance data to improve image processing tasks by generating more accurate and robust algorithms in object detection and tracking [7,8], human activity recognition [9,10], database [11] and tracking performance evaluation tools [12].

The most desirable qualities of a video surveillance system are (a) *robust* operation in real world scenarios, characterized by sudden or gradual changes in the input statistics and (b) *intelligent* analysis of video to assist the operators in scene analysis and event classification. In the past several research works have been carried out in many fields of video surveillance using single vision camera and indeed significant results have been obtained. But mostly they are proven to work in a controlled environment and specific contexts. A typical example is of vehicle and traffic surveillance: systems for queue monitoring, accident detection, car plate recognition etc. In a recent survey on video surveillance and sensor networks research, Cucchiara [13] reports that there are still many unsolved problems in tracking in non ideal conditions, in cluttered and unknown environment, with variable and unfavorable luminance conditions, for

surveillance in indoor and outdoor spaces. Traditional approaches in dealing with these problems have focused on improving the robustness of background model and object segmentation techniques by extracting additional content from data (color, texture etc). However they have used only single modality such as visible spectrum or thermal infrared video. Visible and thermal infrared spectrums are intuitively complementary, since they capture information in emitted and reflected radiations, respectively. Thus alternative approach of integrating information from multiple video modalities has the potential to deal with such dynamically changing environment by leveraging the combined benefits whilst compensating for failures in individual modalities [14].

In addition other media streams like audio can improve analysis of visual data. For example, visual and ambient media capture two different aspects - scene and sound, respectively. In many cases where visual information is not sufficient for reliably discriminating between activities, there is often audio stimulus that is extremely important for a particular classification or anomaly detection task [15].

Automatic intelligent analysis of incoming video data on-line is required because firstly it is practically infeasible to manually supervise huge amount of video data (especially with multiple cameras) and secondly, off-line analysis completely precludes any possibility of taking immediate action in the likely happening of an abnormal event, particularly in critical applications. Several intelligent activity/ event detection methods are being proposed as the behavior patterns of real life scenario still remain challenge for the research community.

Therefore our emphasis in this paper is to discuss the existing (and proposed) techniques and provide summary of progress achieved in the direction of building robust and intelligent surveillance system. The paper's scope goes beyond traditional vision based framework to multi-modal framework. In several places, we briefly review some related concepts in automated surveillance system to put everything in proper context. For detailed discussion on studies in those related areas, reviews are available as follows: Background subtraction techniques [16], tracking of people and body parts [17], face recognition [18], gesture recognition [19], issues in automated visual surveillance [20], multimedia and sensor networks [13], distributed surveillance systems [21] and a detailed review of techniques in all the stages in the general framework of visual surveillance [6].

The rest of the paper is organized as follows. Section 2 gives an overview of automated visual surveillance, its evolution and practical issues. Section 3 discusses computer vision techniques in and beyond visual spectrum that have been developed for object detection and tracking. Section 4 reviews the work related to data fusion in multi modal framework (including visible, infrared and audio). Section 5 covers the activity recognition and behavior understanding approaches for event detection. Finally section 6 concludes the paper by summarizing the discussion and analyzing some possible future research directions.

2 Overview of Automated Visual Surveillance System

The general framework of an automatic video surveillance system is shown in Figure1. Video cameras are connected to a video processing unit to extract high-level information identified with alert situation. This processing unit could be connected throughout a network to a control and visualization center that manages, for example, alerts. Another important component is a video database and retrieval tool where selected video segments, video objects, and related contents can be stored and inquired. In [6, 22], a good description of video object processing in surveillance framework is presented. The main video processing stages include background modeling, object segmentation, object tracking, behaviors and activity analysis. In multi camera scenario, fusion of information is needed, which can take place at any level of processing. Also these cameras may be of different modality like thermal infrared, near infrared, visible color camera etc so that multi spectral video of the same scene can be captured and the redundant information may be used to improve the robustness of the system against dynamic changes in environmental conditions.

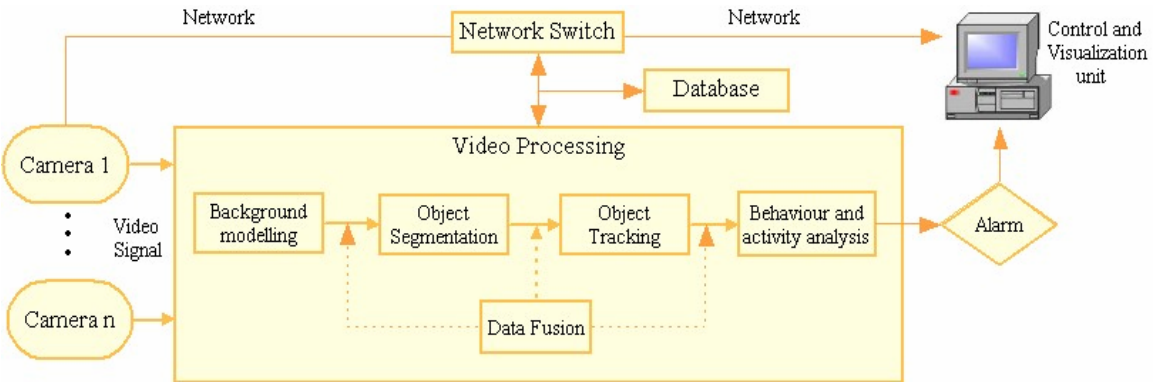

Figure1: General framework of automated visual surveillance system

2.1 Evolution of Surveillance systems

"First generation" video-based surveillance systems started with analog CCTV systems, which consisted of a number of cameras connected to a set of monitors through automated switches. In [23], for example, integration of different CCTV systems to monitor transport systems is discussed. But the human supervision being expensive and ineffective due to widespread deployment of such systems, they are more or less used as a forensic tool to do investigation after the event has taken place. By combining computer vision technology with CCTV systems for automatic processing of images and signals, it becomes possible to proactively detect alarming events rather than passive recording. This led to the development of semi-automatic systems called "second generation" surveillance systems, which require a robust detection and tracking algorithm for behavioral analysis. For example, the real-time visual surveillance system W4 [7] employs a combination of shape analysis and tracking, and constructs models of people's appearances in order to detect and track groups of people as well as monitor their behaviors even in the presence of occlusion and in outdoor environments. Current research issues in such systems are mainly real time robust computer vision algorithms and automatic learning of scene variability and patterns of behaviors.

Third generation surveillance system is aimed towards the design of large distributed and heterogeneous (with fixed, PTZ, and active cameras) surveillance systems for wide area surveillance like monitoring movement of military vehicles on borders, surveillance of public transport etc. For example the Defense Advanced Research Projection Agency (DARPA) supported the Visual Surveillance and Monitoring (VSAM) project [24] in 1997, whose purpose was to develop automatic video understanding technologies that enable a single human operator to monitor behaviors over complex areas such as battlefields and civilian scenes. The usual design approach of these vision systems is to build a wide network of cooperative multiple cameras and sensors to enlarge the field of view. From an image processing point of view, they are based on the distribution of processing capacities over the network and the use of embedded signal processing devices to give the advantages of scalability and robustness potential of distributed systems. The main research problems involved in such systems are: integration of information obtained from different sensors, establishing signal correspondence in space and time, coordination and distribution of processing task and video communication etc.

Recently, the rapid emergence of wireless networks and proliferation of networked digital video cameras have favorably increased the opportunity for deploying large scale Distributed Video Surveillance (DVS) systems on top of existing IP-network infrastructure. Many commercial companies now offer IP-based surveillance solutions. For example companies like Sony and Intel have designed equipments like smart cameras; Cisco provides many networking devices for video surveillance. All this has led to the latest step in the evolution of video-surveillance systems i.e migration to digital IP-based surveillance and recently to wireless interconnection network. Figure 2 shows a general DVS network architecture, where there are several video sensors/cameras distributed over a wide area, with smaller groups under a local base station called Processing proxy server (PPS). A PPS collects video streams from many such video cameras through a wireless (mostly) or wired LAN or mesh network. These servers are equipped with computational power to perform necessary machine vision processing and data filtering to analyze the video stream and identify alert situations. These servers then transmit the video data to different users the backbone internet network.

2.2 Practical issues in Real World Scenario

Despite much advancement in the field, realizing practical an end-to-end video surveillance system in a real world scenario remains a difficult task due to the following issues:

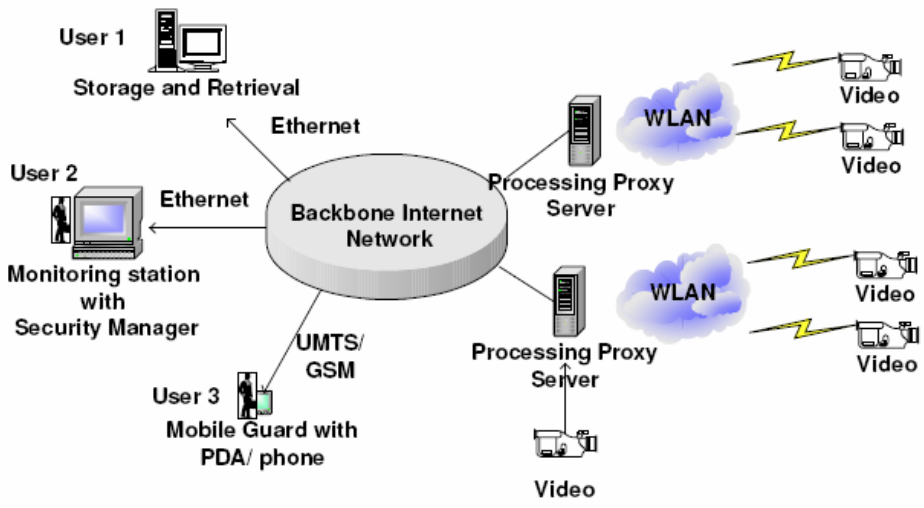

Figure 2: Distributed Video Surveillance Network Architecture

1. Robustness: Real world scenarios are characterized by sudden or gradual changes in the input statistics. A major challenge for real world object detection and tracking is the dynamic nature of real world conditions with respect to illumination, motion, visibility, weather change etc. As pointed out in [22], achieving robust algorithms is a challenge especially (a) under illumination variation due to weather conditions or lighting changes, for example, in outdoor scene, due to movement of clouds in sky and in an indoor scene, due to opening of doors or windows; (b) under view changes; (c) in case of multiple objects with partial or complete occlusion or deformation; (d) in the presence of articulated or non-rigid objects; (e) in case of shadow, reflections, and clutter; and (f) with video noise (e.g., Gaussian white). Figure 3 shows scenarios with (a) low light and illumination variation (b) video noise (c) boat among moving waves and (d) car object with moving background of vegetation. Significant research and advancement in solving these difficulties have been achieved but still the problem is unsolved in generic situation with dynamically varying environmental conditions and there is lack of generic multimodal framework to achieve system robustness by data fusion.

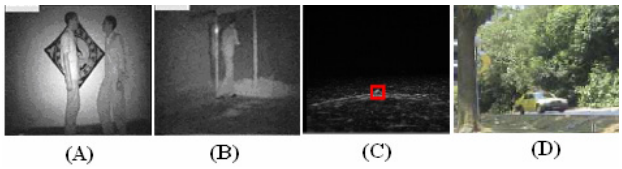

Figure 3: Some examples of complex real world situation for object detection

2. Intelligent: With the advances in sensor technology, surveillance cameras and sound recording systems are already available in banks, hotels, stores and highways, shopping centers and the captured video data are monitored by security guards and stored in archives for forensic evaluation. In a typical system, a security guard watches 16 video channels at the same time and may miss many important events. There is a need of intelligent, (semi-) automated video analysis paradigms to assist the operators in scene analysis and event classification. Event detection is a key component to provide timely warnings to alert security personnel. It deals with mapping motion patterns to semantics (e.g., benign and suspicious events). However detecting semantic events from low-level video features is major challenge in real word situation due to unlimited possibilities of motion patterns and behaviors leading to well known semantic gap issue. Furthermore, suspicious motion events in surveillance videos happen rather infrequently and the limited amount of training data poses additional difficulties in detecting these so-called rare events [25].

3. Real timeliness: A useful processing algorithm for surveillance systems should be real time, i.e., output information's, such as events, as they occur in the real scene [22]. Requirement of accuracy and robustness result in computational intensive and complex design of algorithms which makes real time implementation of system a difficult task.

4. Cost effective: For feasible deployment in a wide variety of real world surveillance applications ranging from indoor intrusion detection to outdoor surveillance of important buildings etc, a cost effective framework is required.

3 Computer Vision techniques for visual surveillance tasks

This section summarizes the research that addresses the basic computer vision problems in video surveillance like object detection and tracking. These modules constitute the low level building block necessary for any surveillance system and we briefly outline the most popular techniques used in these modules. We also present the advances made in computer vision techniques both in and beyond the visible spectrum (thermal infrared etc.) to give motivation for the discussion on data fusion in the next section.

3.1 Object Detection

Nearly every visual surveillance system starts with object detection. Object detection aims at segmenting regions corresponding to moving objects such as vehicles and humans from the rest of an image. Detecting moving regions provides a focus of attention for later processes such as tracking and behavior analysis because only these regions need be considered in the later processes. There are two main conventional approaches to object detection: 'temporal difference' and 'background subtraction'. The first approach consists in the subtraction of two consecutive frames followed by thresholding. The second technique is based on the subtraction of a background or reference model and the current image followed by a labeling process. After applying one of these approaches, morphological operations are typically applied to reduce the noise of the image difference (See figure 4 and 5). The temporal difference technique is very adaptive to changes in dynamic environment and another advantage is that it does not make assumptions about the scene. However, it can be problematic that only motion at edges is visible for homogeneous objects. On the other hand, background subtraction has better performance extracting object information but it is sensitive to dynamic changes in the environment.

Figure 4: Example of object detection using temporal differencing technique

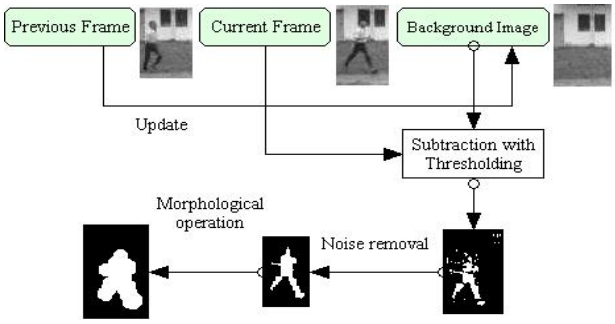

Figure 5: Object Detection using background subtraction technique

Background modeling assumes that the video scene is composed of a relatively static model of the background, which becomes partially occluded by objects that enter the scene. These objects (usually people or vehicles) are assumed to have features that differ significantly from those of the background model (their color or edge features, for example). The terms *foreground* and *background* are not scientifically defined however and thus their meaning may vary across applications. For example, a moving car should usually be considered as a foreground object but when it parks and remains still for a long period of time, it is expected to become background. Also, not all moving objects can be considered foreground. The simplest approach is to record an image when no objects are present and use this image as the background model. However, continuous updating of the model is required to make the foreground extraction more robust to the gradual changes in lighting and movement of static objects that are to be expected in outdoor scenes. Unfavorable factors, such as illumination variance, shadows and shaking branches, bring many difficulties to the acquirement and updating of background model. Background modeling is a very active research area and several techniques have been proposed to deal with various problems. A good overview of the most frequently cited background modeling algorithms is given in [16]. A comparison between various background modeling algorithms is given in [26], as well as a discussion on the general principles of background maintenance systems.

A typical approach for modeling background in outdoor conditions is using Gaussian model that models the intensity of each pixel with a single Gaussian distribution [27] or with more than one Gaussian distribution. The algorithm described in [28] models each pixel as a sum of K Gaussian distributions in RGB space ($1 \leq K \leq 5$). Each pixel's background model is updated continuously, using online estimation of the parameters. This model is well suited to cater for pixels whose background model has a multimodal distribution, such as vegetation or water. The model is unable to distinguish between foreground objects and shadows, however, and also is quite slow to initialize. The algorithm used in the $W4$ [6] system works on monochrome video and marks a pixel as foreground if:

$$|M - It| > D \text{ or } |N - It| > D \text{ (1)} \quad (1)$$

where the (per pixel) parameters M, N, and D represent the minimum, maximum, and largest inter-frame absolute difference observed in the training frames. It also detects when its background model is invalid by detecting when 80% of the image appears as foreground. To rectify this, is re-enters a training mode to correct the background model. However reliable background modeling is difficult to achieve in certain scenarios. For example, in a crowded room with many people, the background may only ever be partially visible. Another problematic scenario is in a scene with low levels of lighting, such as a night-time scene with only street lighting. The movement (or apparent movement) of background objects is problematic too. Examples of this include moving trees and vegetation, flickering computer or TV screens, flags or banners blowing in the wind, etc.

Apart from common approached discussed above, techniques based on optic flow is useful in motion segmentation where a motion vector is assigned to every pixel of the image by comparison of successive frames. Optical-flow-based methods can be used to detect independently moving objects even in the presence of camera motion. However, most flow computation methods are computationally complex and very sensitive to noise, and cannot be applied to video streams in real time without specialized hardware. More detailed discussion of optical flow can be found in Barron's work [29].

3.2 Object Tracking

Once objects have been detected, the next logical step is to track these detected objects. Tracking has a number of benefits. Firstly, the detection phase is quite computationally expensive, so by using tracking, the detection step does not need to be computed for each frame. Secondly, tracking adds temporal consistency to sequence analysis because otherwise, objects may appear and disappear in consecutive frames due to detection failure. Also, tracking can incorporate validity checking to remove false positives from the detection phase. Thirdly, if tracking multiple objects, detection of occlusion is made easier, as we expect occlusion when two or more tracked object move past each other (as shown in figure 6). Object motion can be perceived as a result of either camera motion with a static object, object motion with static camera, or both object and camera moving. Tracking techniques can be divided into two main approaches: 2-D models with or without explicit shape models and 3-D models. For example, in [30] the 3-D geometrical models of a car, a van and a lorry are used to track vehicles on a highway. The model-based approach uses explicit a priori geometrical knowledge of the objects to follow, which in surveillance applications are usually people, vehicles or both. In [6], author use a combination of shape analysis and along with 2D Cardboard Model for representing and tracking the different body parts. Along with second order predictive motion models of the body and its parts, they used

Cardboard Model to predict the positions of the individual body parts from frame to frame.

A common tracking method is to use a filtering mechanism to predict each movement of the recognized object. The filter most commonly used in surveillance systems is the Kalman filter [31]. Fitting bounding boxes or ellipses, which are commonly called 'blobs', to image regions of maximum probability is another tracking approach based on statistical models. In [27] the author models and tracks different parts of a human body using blobs, which are described in statistical terms by a spatial and color Gaussian distribution. In some situations of interest the assumptions made to apply linear or Gaussian filters do not hold, and then nonlinear Bayesian filters, such as extended Kalman filters (EKF) or particle filters have been proposed. A good tutorial on non linear tracking using particle filter is given in [32] where the author illustrates that in highly non-linear environments particle filters give better performance than EKF. A particle filter is a numerical method, which weights (or 'particle') a representation of posterior probability densities by resampling a set of random samples associated with a weight and computing the estimate probabilities based on these weights. Then, the critical design decision using particle filters relies on the choice of importance (the initial weight) of the density function. Appearance models [33] are another way to represent objects. It consists of an observation model (usually an image) of the tracked object, along with some statistical properties (such as the pixel variances).

Figure 6: Object Tracking during and after Occlusion

3.3 Computer Vision beyond the Visible Spectrum

Traditionally, the majority of the computer vision community has been involved implicitly or explicitly with the development of algorithms associated with sensors that operate in the visible band of the electromagnetic spectrum [34]. In the past, imaging sensors beyond visible spectrum have been limited to special applications like remote sensing and vision based military applications, because of their high cost. Recently with the advances in sensor technologies, the cost of near and mid-infrared sensors has dropped dramatically, making it feasible for their use in more common applications like automatic video-based security and surveillance systems to enhance their capabilities.

Lin, in his 2001 technical report explores the extension of visible band computer vision techniques to infrared as well as conducts a good review of infrared imaging research [35]. Recent literature on the exploitation of near-infrared information to track humans generally deals only with the face of observed people[36] and a few are concerned with the whole body [37,38] but these approach rely on the highly limiting assumption that the person region always has a much brighter (hotter) appearance than the background. This assumption does not hold in various weather conditions and during all time. To tackle this, the author in [39] proposes a novel contour based background subtraction strategy to detect people in thermal imagery, which is robust across a wide range of environmental conditions. First of all, a standard background-subtraction technique is used to identify local region-of interest (ROI), each containing the person and surrounding thermal halo. The foreground and background gradient information within each region are then combined into a contour saliency map (highlighting the person boundary). Using a watershed-based algorithm, the gradients are thinned and thresholded into contour fragments. The remaining watershed lines are used as a guide for an A* search algorithm to connect any contour gaps. Finally, the closed contours are flood-filled to make silhouettes.

The use of infrared in pedestrian detection to reduce night time accidents is investigated in [38]. In [40], the author investigates human repetitive activity properties using thermal imagery. They employ a spatio-temporal representation, Gait Energy Image (GEI), which represent human motion sequence in a single image while preserving some temporal information. However they have developed the method for only simple activities which are repetitive in nature (like walking, running etc).

4 Data Fusion

A surveillance task using multiple modalities can be divided into two major phases: *data fusion* and *event recognition*. The data-fusion phase integrates multi-source spatio-temporal data to detect and extract motion trajectories from video sources. The event-recognition phase deals with classifying the events as to relevance for the search. This section discusses the data fusion part and the event-recognition task is discussed in the next section.

Data fusion is the process of combining data from multiple sources such that the resulting entity or decision is in some sense better than that provided by any of the individual sources. Most of the existing surveillance systems have used only one media (i.e. normal video), and therefore they do not capture different aspects of the environment. Multiple media are useful because each media captures different aspect of the environment. For example sensing environmental sound can provide reliable clue for detecting insecure events in many cases. Infrared is more informative in dark environment, especially at night. Visible and thermal infrared spectrums are intuitively complementary, since they capture information in emitted and reflected radiations, respectively. Thus combining them can be advantageous in many scenarios, especially when one modality perform poorly in detecting objects. For example visible analysis has an obvious limitation of daytime operation only and completely fails in total darkness. Additionally foggy weather condition, sudden lighting changes,

shadows and color camouflage, often cause poor segmentation of actual objects and much false positive detection. Thermal infrared video is almost completely immune to lighting changes, and thus it is very robust to the above mentioned problems. However, infrared video has its unique inherent challenges due to high noise, and "Halo effect" produced by some infrared sensors, which appears as a dark or bright halo surrounding very hot or cold objects respectively. Further if people are wearing insulated clothing or infrared camera performs rapid automatic gain then it will cause foreground detection to incorrectly classify pixels. See figure 7 for illustration in two different situations. Thus by data fusion approach, it is possible to improve robustness of the system in dynamic real world conditions.

4.1 Fusion of Visible and Infrared

Depending on the application and fusion method, research in the fusion of visible and infrared imagery can be classified in two broad categories. *Image based Representational fusion* and *Video based Analytical fusion*. In *Image based Representational fusion,* the goal is to obtain best representation of the data in a single image for improved visual perception by combining multiple images to create a single fused image that somehow represents the information content of the input images. This type of fusion is generally used for remote sensing and military applications. Depending on the synergy of the information inherent in the data, it may be possible to reduce noise, to extend the field of view beyond that of any single image, to restore high frequency content, and even to increase spatial resolution [41]. Image fusion techniques have had a long history in vision. Gradient-based techniques examining gradients at multiple resolutions [42] and several region-based multi resolution algorithms have been proposed such as the pyramid approaches of [43, 44] and the wavelet-based approach of [45].

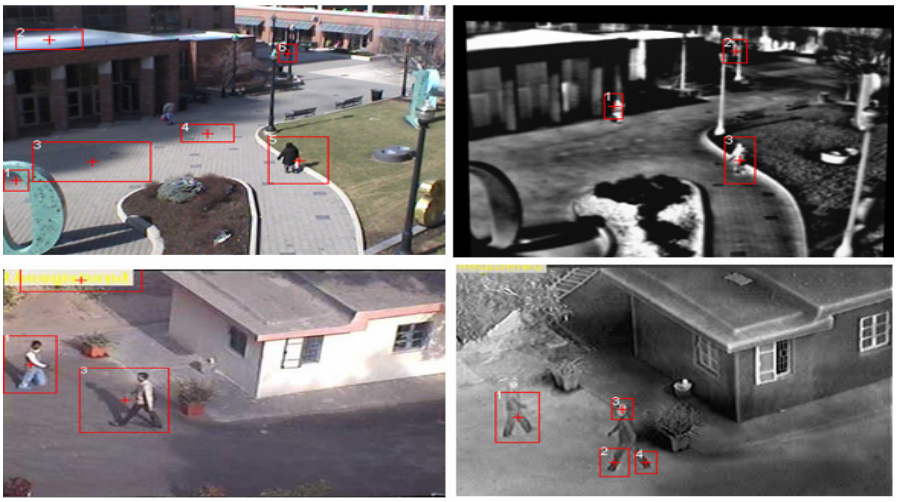

Figure 7: Visible and corresponding thermal infrared in a) Variable illumination due to cloud movement causing false detection in visible (top) b) Incorrect detection due to shadows in visible and thermally insulated clothing in infrared (bottom)

Video based Analytical Fusion, on the other hand, aims to extract knowledge by using all sources of data for better analysis, and not merely to represent the data in another way. This type of fusion methodology is required to enhance the capabilities of automatic video-based detection and tracking system for surveillance purpose. Although image fusion has received considerable attention in the past, research in the fusion of video modalities for automatic analysis, or analytical fusion is very recent. Some recent works have addressed the tracking of humans and vehicles with multiple sensors [46, 47] but issues that are involve in fusing multiple modalities for robust detection and tracking is very sparse. In [48], the fusion of thermal infrared with visible spectrum video, in the context of surveillance and security, is done at the object level. Detection and tracking of blobs (regions) are performed separately in the visible and thermal modality. An object is made up of one of more blobs, which are inherited or removed as time passes. Correspondences are obtained between objects in each modality, forming a master-slave relationship, so that the *master* (the object with the better detection or *confidence*) assists the tracking of the slave in the other modality. Their system uses many heuristics and there also seems to be many parameters to set empirically.

Davis et al. [49] propose a new contour-based background-subtraction technique using thermal and visible imagery for persistent object detection in urban settings. They perform statistical background subtraction in the thermal domain to identify the initial regions-of-interest. Color and intensity information are used within these areas to obtain the corresponding regions of-interest in the visible domain. Within each image region (thermal and visible treated independently), the input and background gradient information are combined as to highlight only the boundaries of the foreground object. The boundaries are then thinned and thresholded to form binary contour fragments. Contour fragments belonging to corresponding regions in the thermal and visible

domains are then fused using the combined input gradient information from both sensors. An A* search algorithm constrained to a local watershed segmentation is then used to complete and close any contour fragments. Finally, the contours are flood-filled to make silhouettes.

In a very recent work [50], the authors use thermal infrared video with standard CCTV video for object segmentation and retrieval in surveillance video. They segment object using separate background modeling in each modality and dynamic mutual fusion based thresholding. Transferable Belief Model is used to combine the sources of information for validating the tracking of objects. Extracted objects are subsequently tracked using adaptive thermo-visual appearance. However they don't take into account the reliability of each source in the fusion process. In [51], an intelligent fusion approach using Fuzzy logic and Kalman filtering technique is discussed to track objects and obtain fused estimate according to the reliability of the sensors. Appropriate measurement parameters are identified to determine the measurement accuracy of each sensor. A comparison of multiple fusion schemes for appearance based tracking of objects using thermal infrared and visible modalities is done in [52] for different objects, such as people, faces, bicycles and vehicles.

4.2 Data Fusion Methods

For visual surveillance using multiple cameras, issues such as camera calibration and registration, establishing correspondences between the objects in different image sequences taken by different cameras, target tracking and data fusion need to be addressed. The success of information fusion depends on how well data are represented, how reliable and adequate the model of data uncertainty used and how accurate and applicable prior knowledge is. Three commonly used fusion approaches are probabilistic methods (Bayesian inference), fuzzy logic method and belief models (Dempster-Shafer model and Transferable Belief model). The Bayesian inference method quantitatively computes the probability that an observation can be attributed to a given assumed hypothesis but lacks in ability to handle mutually exclusive hypotheses and general uncertainty [53]. Fuzzy logic methods accommodate imprecise states and variables. It provides tools to deal with observations that is not easily separated into discrete segments and is difficult to model with conventional mathematical or rule-based schemes [54]. The Belief theory generalizes Bayesian theory to relax the Bayesian method's restriction on mutually exclusive hypotheses, so that it is able to assign evidence to 'propositions', i.e. unions of hypotheses. Dempster-Shafer model makes a closed world assumption, so it assigns a belief of empty set to zero. The reasoning model assumes completeness of the frame of discernment meaning that the frame includes all hypotheses. But it can very well happen that some hypotheses, because of measurements are excluded from frame of discernment or unknown. In this way meaning of empty set is changed corresponding not only for impossibilities but also for unknown possibilities [55]. This kind of approach is called open world assumption which is considered in another belief model called Transferable Belief Model (TBM) [56]. TBM offers the flexibility to model closed world or open world assumption.

In [57], Bayesian probability theory is used to fuse the tracking information available from a suite of cues to track a person in 3D space. In [58], the authors uses TBM framework to solve the problem of data association in a multi target detection problem. It uses the basic belief mass $m(\Theta)$ as a measure of conflict and the sensors are clustered so that the conflict is minimized. But they tackle only partial problem of assessing how many objects are present and observed by the sensors. In [59], the author use TBM and Kalman filter for data fusion in object recognition system that analyses simulated FLIR and LADAR data to recognize and track aircraft. They demonstrated the results on an air to air missile based simulation system. In [60], the author proposes a hybrid multi-sensor data fusion architecture using Kalman filtering and fuzzy logic techniques. They feed the measurement coming from each sensor to separate fuzzy–adaptive kalman filters (FKF), working in parallel. The adaptation in each FKF is in the sense of adaptively adjusting the measurement noise coar1ance matrix R employing a fuzzy inference system (FIS) based on a covariance matching technique. Another FIS, which they call as fuzzy logic observer (FLO) monitors the performance of each FKF. Based on the value of a variable called Degree of Matching *(D0M)* and the matrix R coming from each FKF, the FLO assigns a degree of confidence, a number on the interval (0, 1], to each one of the FKFs output. The degree of confidence indicates to what level each FKF output reflects the true value of the measurement. Finally, a defuzzificator obtains the fused estimated measurement based on the confidence values. They demonstrated the result theoretically, by taking example of four noisy inputs.

4.3 Reliability of Sensor

In the fusion process, different sources may have different reliability and it is essential to account for this fact to avoid decreasing in performance of fusion results. The fused estimate should be more biased by accurate measurements and almost unaffected by inaccurate or malfunctioning ones. Therefore for fusing data collected from different sensors requires the determination of measurements' accuracy so that they can be fused in a weighted manner. The most natural way to deal with this problem is to establish reliability of the beliefs computed within the framework of the model selected. For example [61] discusses a method for assessing the reliability of a sensor in a classification problem within the TBM framework. In [62], the authors propose a multi-sensor data fusion method for video surveillance, and demonstrated the results by using optical and infrared sensors. The measurements coming from different sensors were weighted by adjusting measurement error covariance matrix used by the fusion filter. To estimate

the reliability of the sensor they defined a metric called Appearance Ratio (AR), whose value is proportional to the strength of the segmented blobs from each sensor. The ARs are compared to determine which sensors are more informative and therefore selected to perform a specific video surveillance task. In [63], the authors discuss the principal concepts and strategies of incorporating reliability into classical fusion operators and provide good literature survey on main approaches used in fusion literatures to estimate reliability of sensor.

4.4 Audio and Video Information Fusion

Enhancing visual data with audio streams can serve manifold purpose like speaker tracking, environment sound recognition for event recognition in surveillance application etc. Environmental sound like that of breaking of glass, dog's barking, screaming of a person, fire alarm, gun firing and similar kind of sounds, if detected and recognized correctly, can give a reasonable degree of confidence in making a decision about 'secure' vs. 'insecure' state [64]. Multimedia researchers have often used early fusion strategy to perform the audio-visual fusion for various problems including speech processing [65] and recognition [66], speaker localization [67] and tracking [68, 69], and monologue detection [70]. In [68], the authors present a method that fuses 2-D object shape and audio information via importance filters. They used audio information to generate an importance sampling function, which guides the random search process of particle filter towards regions of the configuration space likely to contain the true configuration (a speaker). A recent work in [71], describes a process to assimilate data from coarse and medium grain sensors, namely video and audio, and a probabilistic framework to discriminate concurring and contradictory evidences. The authors enlarge the concept to information fusion with the definition of *information assimilation*: this process includes not only the real-time information fusion but also the integration with the past experience, represented by the surveillance information stored in the system.

Research in the field of environmental sound recognition is sparse. The majority of auditory research is centered on the identification and recognition of speech signals. Those systems that do exist, work on a very specialized domain like in [72], a system named *AutoAlert* is presented for automated detection of incidents using HMM, and Canonical Variates Analysis (CVA) to analyze both short-term and time-varying signals that characterize incidents. Cowling, in [73] provides more detailed literature survey, and it also investigates few existing techniques used for sound recognition in speech and music. He then presents a comparison on the accuracy of these techniques, when employed for the problem of non-speech environmental sound classification for autonomous surveillance.

4.5 Other Sensors/modalities

There are proximity sensors (like ultrasound devices, lasers scanners etc.) which detect objects without physical contact. Most proximity sensors emit an electromagnetic field or beam and look for changes in the field. Different targets demand different sensors. For example, a capacitive or photoelectric sensor might be suitable for a plastic target; an inductive sensor requires a metal target [74]. In [75], the authors integrate Laser Doppler Vibrometer (LDV) and IR video for remote multimodal surveillance. Their work mainly caters to remote area surveillance and their main focus was to study the application of LDV for remote voice detection, while IR imaging was used for target selection and localization. Few object tracking and visual servoing system for the visually impaired such as the *GuideCane* [76] and the *NavBelt* [77], use ultrasound or laser rangefinders to detect obstacles.

Gated imaging is another useful system for highly unfavorable visual conditions like underwater surveillance etc. Time gating is a temporal example of image formation whereby a light source is time pulse projected toward a target and the detector is time gated to accept image-forming illumination from a specific range [78]. LIDAR systems [79] time gate the receiver aperture to eliminate relatively intense backscatter originating from the water while allowing the return from the target to be detected. In [78] time gating is employed for using spatially and temporally varying coherent illumination for undersea object detection.

5 Event Detection

5.1 Human Activity Recognition

Computer analysis of human actions is gaining increasing interests, especially in video surveillance arenas where people identification and activity recognition are important. Using two important metrics: preciseness of the analysis outcome and the required video resolution to achieve the desired outcome, human identification and activity recognition can be classified into three categories. At one extreme, which is often characterized by high video resolution and a small amount of scene clutter, high fidelity outcome is achievable. Many techniques in face, gesture, and gait recognitions fall in this category, which aim to identify individuals against a pre-established database.

At the other extreme, which is characterized by low video resolution and potentially significant scene clutter, it is often not possible to achieve highly discriminative outcome. Instead, the goal is often to detect the presence, and identify the movement and interaction of people through "blob" tracking [6, 24]. The VSAM system [24] tracked the human body as a whole blob. They use a hybrid algorithm by combining adaptive background subtraction with a three-frame differencing technique to detect moving objects, and use the Kalman filter to track the moving objects over time. A neural network classifier is trained to recognize four classes: single person, group of persons, vehicles, and clutter. They also use linear discriminant analysis to further provide a finer distinction between vehicle types and colors. The VSAM system is

very successful at tracking humans and cars, and at discriminating between vehicle types. But it did not put much emphasis on activity recognition; only gait analysis and simple human-vehicle activity recognition are handled.

In the middle of the spectrum, it is possible to refine the "blob" representation of a person through hierarchical, articulated models. For example, [80] describes an approach which attempts to recognize more generic activities and movement of body parts using MHI (motion-history images) to record both the segmentation result and the temporal motion information. The MHI is a *single* image composed of superimposing a sequence of segmented moving objects weighed by time. The most recent foreground pixels are assigned the brightest color while past foreground pixels are progressively dimmed. This allows the summarization of information on both the spatial coverage and the temporal ordering of the coverage of an activity. See figure 8 for illustration with few examples. The MHI does not use any structure to model human. A vector of seven moment values is computed for each MHI. Activities are recognized by finding the best match of the moment vectors between the query MHI and the training patterns. Other approaches allow main body parts, such as head, arms, torso, and legs, to be individually identified to specify the activities more precisely. A very detailed review of activity recognition approaches in these categories can be found in [8].

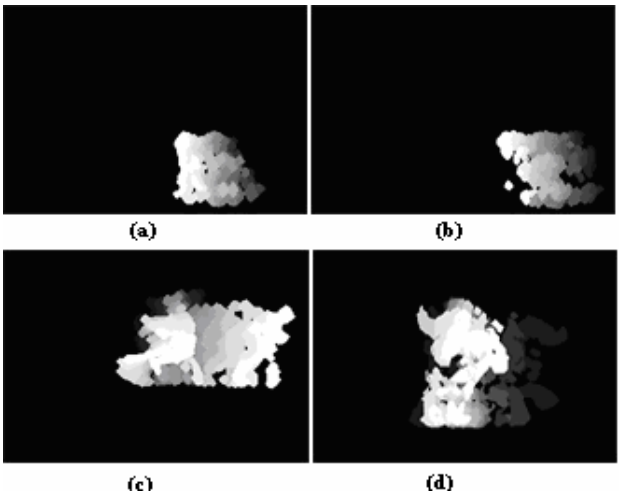

Figure 8: MHI images of person (a) walking (b) Running, (c) picking an object and (d) fighting

5.2 Semantic Information Extraction for Behavior Understanding

Understanding of behaviors may simply be thought as the classification of time varying feature data, i.e., to analyze the video to extract some feature vectors and to classify this time-varying feature data. During the recognition phase, extracted unknown test feature vector set is compared to a group of labeled reference feature vector sets representing typical human actions.

Feature extraction process is very important to achieve good results. It is impossible to train and perform classification using the currently available classification engines. The size of the feature vector should be as small as possible to have computational efficiency. At the same time it should represent each action very accurately. For example, the center of mass of the tracked object in image frame of the video can be used as a feature vector in a security application in which moving persons entering or leaving a building. In this simple problem, the "vocabulary" consists of a person leaving the building and a person entering a building and the center of mass information consisting of the horizontal and vertical coordinates in an image of the video may be good feature vector for this problem. On the other hand, detecting an 'assault' case where a person falls on the ground and other runs will require other additional parameters in the feature set. For example, so-called snaxels of an active snake contour of the human body can be added to the feature vector to distinguish a fallen person from a person standing. Compactness of the contour boundary, the speed of the center of the mass etc can be also used to distinguish the normal action of walking and the abnormal action of a falling and running. As a rule of thumb, model parameters are selected as entries of the feature vector in a model based tracking approach. Since a video consists of sequence of images a sequence of feature vectors are obtained to characterize the motion of a person(s).

5.3 Pattern Analysis and Classification Methods

Several generative and discriminative models have been proposed for modeling and classifying activity patterns. Some of the most widely used ones are Dynamic time Warping (DTW), Hidden Markov Models (HMM), Time Delay Neural Network (TDNN), and Finite State Machine (FSM) network.

a) Dynamic time warping: DTW is a template based dynamic programming matching technique widely used in the algorithms for speech recognition. It has the advantage of conceptual simplicity and robust performance, and has been used recently in the matching of human movement patterns [81]. For instance, Bobick *et al.* [82] use DTW to match a test sequence to a deterministic sequence of states to recognize human gestures. Even if the time scale between a test sequence and a reference sequence is inconsistent, DTW can still successfully establish matching as long as the time ordering constraints hold.

b) Hidden Markov Models: HMMs are stochastic finite state machines [83]. In the context of human motion analysis, a finite state Markov model is assigned for each possible scenario and its parameters are trained with feature vectors of this typical human action. The training process is an off-line iterative algorithm called Baum-Welch algorithm [84]. Here , the number of states of a HMM must be specified, and the corresponding state transition and output probabilities are optimized in order that the generated symbols can correspond to the

observed image features of the examples within a specific movement class. During the classification or recognition phase, the test feature vector set is applied to all of the Markov models and output probabilities are computed. The Markov model producing the highest probability is determined and the corresponding human action scenario is selected as the result. HMMs generally outperform DTW for undivided time series data, and are therefore extensively applied to behavior understanding. In [85], authors describe an activity recognition process for visual surveillance of wide areas and experimented with image sequences acquired from an archaeological site with actors perform both legal and illegal actions. The activity recognition process is performed in three steps: first of all the binary shape of moving people are segmented, then the human body posture is estimated frame by frame and finally, for each activity to be recognized, a temporal model of the detected postures is generated by Discrete Hidden Markov Models

c) Finite state machine: The most important feature of a FSM is its state-transition function. The states are used to decide which reference sequence matches with the test sequence. However it requires hand crafted heuristic rules based on context knowledge. For example in [86], the authors propose a framework for unsupervised learning of usual activity patterns and detection of unusual activities based on a model of multi-layered finite state machines. They considered two different approaches for different scenario. First approach is unsupervised learning of usual activity patterns and detection of unusual activities (which are not recognized as normal). Other approach is to explicitly program the FSM or training using supervised learning for recognition of situation specific activities like unattended baggage detection.

d) *Time-delay neural network (TDNN):* TDNN is also an interesting approach to analyzing time-varying data. In TDNN, delay units are added to a general static network, and some of the preceding values in a time-varying sequence are used to predict the next value. As larger data sets become available, more emphasis is being placed on neural networks for representing temporal information. TDNN has been successfully applied to hand gesture recognition [87] and lip-reading [88].

Other related schemes including Dynamic Bayesian Network (DBN) and the Support Vector Machines (SVM) are also now being actively used in pattern analysis and classification problems.

Considering the limited training data for unusual events and that the distinction between two unusual events can be as large as those between unusual events and usual events, it is not feasible to train a general model for the unusual events. Therefore alternative approach that some people have taken is to train a model for usual events and events that deviate significantly from the usual event model are considered unusual. For example a simple ATM surveillance scenario is shown in figure 9, where FSM can be a useful way to discriminate between normal and abnormal patterns. The first row shows frames corresponding to normal transaction. Second and third row shows event corresponding to vandalism and robbery.

Figure 9: Sampleshots for events in a Bank ATM Surveillance

Figure 10 shows a possible FSM for activities which will be considered normal if the transitions terminated at the exit node. If exit is made by another node due to any deviant pattern, the FSM will flag an abnormal event.

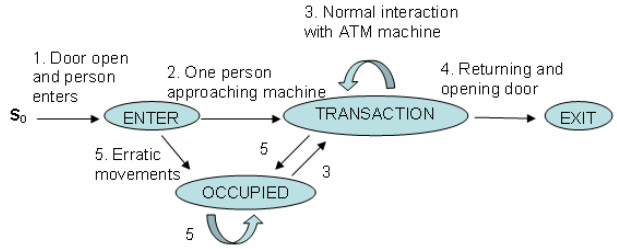

Figure 10: A possible FSM for Bank ATM Surveillance

Boiman and Irani [89] proposed to model the set of usual events as an ensemble of spatial-temporal image patches, and detect irregularity in a test video by evaluating the similarity between the test ensemble with the training ensemble. Zhang *et al.* [90] used a semi-supervised approach to train models for both usual and unusual events. They start from an ergodic hidden Markov model (HMM) for usual events. If a test event does not fit the model, they classify it as unusual and branch the usual event model to refit the usual event. This approach has a disadvantage that it may give high false positive alarm because in a practical scenario, even during the normal course of activity, unusual deviations are very likely to happen without potential threat.

6 Conclusions and Future Research Developments

Event though significant progress have been made 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 are compounded by practical

considerations such as robustness to unfavorable weather and lighting conditions, intelligent video processing for event detection and efficiency in terms of real time operation and cost. Most surveillance systems operate using single modality and lack robustness because they are limited to a particular situation. Different media source like audio, video and thermal infrared gives complementary/supplementary surveillance information of the environment. The literature survey on multi-modal data fusion shows that effective fusion of the information coming from different media streams can give robustness to real world object detection, tracking and event detection. Simultaneous fusion of thermal infrared and visible spectrum can give following advantage:

- Improved robustness against camouflage as foreground object are less likely to be of similar *color and temp* to the background
- Providing features that can be used for classification or retrieval of objects in large surveillance video archives
- Extracting a signature of an object in each modality, which indicates how useful each modality is in tracking that object

However, the key challenge for future research is to: 1) develop analysis techniques to automatically determine the reliability of data source and 2) develop a suitable fusion methodology that would intelligently utilize the information provided by these two modalities to get the best possible output. In [91], these challenges are addressed by employing Transferable Belief Model (TBM) and Kalman filter. TBM is used to determine the validity of a foreground region, detected by each source, for tracking. Kalman filter is used for the dual purpose of tracking the objects over time and fusing the measurements of the positions of the target obtained from different sensors, according to their reliability.

One of the main objectives of visual surveillance is to analyze and interpret individual behaviors and interactions between objects for event detection. Recently, related research has still focused on some basic problems like recognition of standard gestures and simple behaviors. Some progress has been made in building the statistical models of human behaviors using machine learning. However behavior recognition is complex, as the same behavior may have several different meanings depending upon the scene and task context in which it is performed. An alternative approach can be of providing *selective focus-of-attention* to the human supervisor by discriminating unusual or anomalous event from normal ones. Use of audio is encouraging in this respect because in many cases when the visual information required for a particular task is extremely subtle, audio stimulus is extremely salient for a particular anomaly detection task. Moreover, audio features can help in mining interesting patterns from the scene. Multimedia data mining has been applied for detection of events in sports video (like goal event in soccer [25] etc.) but it has not been systematically applied for surveillance videos. This requires development of a novel framework of low level feature extraction, advanced temporal analysis and multimodal data mining methods.

An obvious requirement for surveillance system is real time performance. If the system has to process signals from multiple sensor and modalities, then the required processing is multiplied. Moreover requirement for robustness and accuracy tend to make the algorithm design complex and highly computational. Generally commercial products employ embedded signal processing devices and high performance dedicated processors for faster processing but simultaneously increase the system cost heavily. Past research have focused on optimizing the low level image processing algorithms, reducing the feature space etc and recently researches on distributed surveillance system have tried to distribute the processing on the network. However there is lack of any systematic design and real time implementation of video processing algorithms using a network of multiple processors. Recent research and development in Grid technology can provide an alternative architecture for such implementation and research needs to be done to explore the feasibility of such innovative approach.

References

[1] DataMonitor. Global digital video surveillance markets: Finding future opportunities as analog makes way for digital. Market research report (July 2004). www.mindbranch.com/products/R313-6950.html.

[2] T. Bodsky, R. Cohen, E. Cohen-Solal, S. Gutta, D. Lyons, V. Philomin, and M. Trajkovic (2001). 'Visual surveillance in retail stores and in the home', *in: 'Advanced Video-based Surveillance Systems'*. Kluwer Academic Publishers, Boston, , Chapter 4, pp. 50–61

[3] J.M. Ferryman, S.J. Maybank, and A.D. Worrall (2000). 'Visual surveillance for moving vehicles', *Int. J. Comput. Vis.*, 37, (2), Kluwer Academic Publishers, Netherlands, pp. 187–19731

[4] R. Cucchiara, C. Grana, A. Patri, G. Tardini, and R. Vezzani (2004). Using computer vision techniques for dangerous situation detection in domotic applications. *Proc. IEE Workshop on Intelligent Distributed Surveillance Systems,* London, pp. 1–5.

[5] Frost & Sullivan report, (August 2005); www.frost.com

[6] W. Hu, T. Tan, L. Wang, and S. Maybank (August 2004). A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man and cybernetics*, 34(3):334–350,

[7] I. Haritaoglu, D. Harwood, and L. Davis. W4: Who, when, where, what: A real time system for detecting and tracking people. *In Third Face and Gesture Recognition Conference*, pp. 222-227.

[8] W. Niu, L. Jiao, D. Han, and Y. Wang (2003). Real-Time Multi-Person Tracking in Video Surveillance. *Proceedings of the Pacific Rim Multimedia Conference*, Singapore.

[9] A. Bobick and J. Davis (December 1996). Real time recognition of activity using temporal templates. IEEE workshop on Applications of Computer Vision, Sarasota, FL, 4 pages.

[10] J. Gao, A. G. Hauptmann and H. D. Wactlar. Combining Motion Segmentation with Tracking for Activity Analysis (2004). *The Sixth International Conference on Automatic Face and Gesture Recognition* (FGR'04), pp. 699-704, Seoul, Korea, May 17-19.

[11] E. Stringa,, and C.S. Regazzoni. Content-based retrieval and real-time detection from video sequences acquired by surveillance systems. *Int. Conf. on Image Processing, Chicago*, 1998, pp. 138–142

[12] J. Black, T. Ellis and P. Rosin. A novel method for video tracking performance evaluation (2003). *The Joint IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, October, France, pp. 125–132.

[13] R. Cucchiara (2005). Multimedia surveillance systems. In *VSSN 05: Proceedings of the third ACM international workshop on Video surveillance & sensor networks, New York NY, USA*, pages 3-10.

[14] C. ´O Conaire, E. Cooke, N. O'Connor, N. Murphy, and A. F.Smeaton (2005). Fusion of infrared and visible spectrum video for indoor surveillance. *In International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS),* Montreux, Switzerland, April.

[15] C. stauffer. Automated Audio-Visual Activity Analysis. *CSAIL Technical Reports*, MIT. http://hdl.handle.net/1721.1/30568

[16] A. M. McIvor (2000). Background subtraction techniques. In *Image and Vision Computing, Hamilton, New Zealand*, Nov.

[17] J. J. Wang and S. Singh (2003). Video analysis of human dynamics – a survey. *Real-Time Imaging 9(5):* 321-346.

[18] R. Chellappa, C. L. Wilson and S. Sirohey (1995). Human and machine recognition of faces: a survey. *Proc. of the IEEE*, vol. 83, No. 5, pp705-740.

[19] V. I. Pavlovic, R. Sharma and T. S. Huang (1997). Visual interpretation of hand gestures for human computer interaction: a review. *IEEE Transactions on PAMI*, vol.19, no.7, pp.677-695, July.

[20] A. R. Dick and M. J. Brooks (2003). Issues in Automated Visual Surveillance. DICTA: 195-204.

[21] M. Valera and S.A. Velastin (2005). Intelligent distributed surveillance systems: a review vision. *In Image and Signal Processing, IEE Proceedings*, volume 152, pages 192 – 204, April.

[22] A. Amer and C. Regazzoni. *Editorial:* Introduction to the special issue on video object processing for surveillance applications. *Real-Time Imaging* Vol. 11, pp: 167–171, 2005.

[23] Nwagboso, C (1998). User focused surveillance systems integration for intelligent transport systems', *in 'Advanced Video-based Surveillance Systems'*. Kluwer Academic Publishers, Boston, Chapter 1.1, pp. 8–12.

[24] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y.Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, and L.Wixson (2000). A system for video surveillance and monitoring. *Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep*., CMU-RI-TR-00-12.

[25] M. Chen, S-C. Chen, M-L. Shyu, and K. Wickramaratna. Semantic Event Detection via Multimodal Data Mining. *IEEE Signal Processing Magazine*. pages 38-46, march 2006.

[26] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers (1999). Wallflower: Principles and practice of background maintenance. In *Proceedings of the Seventh IEEE International Conference on Computer IEEE Comput. Soc.*, volume 1, pages 255–261.

[27] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland (1997). Pfinder: real-time tracking of the human body. *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 780–785, July.

[28] C. Stauffer and W.E.L. Grimson (1999). Adaptive background mixture models for real-time tracking. In *Proceedings of CVPR99*, pages II:246–252.

[29] J. Barron, D. Fleet, and S. Beauchemin (1994) .Performance of optical flow techniques. *Int. J. Comput.Vis.*, vol. 12, no. 1, pp. 42–77.

[30] J.M. Ferryman, S.J. Maybank, and A.D. Worrall (2000). Visual surveillance for moving vehicles, Int. J. Comput. Vis., 37, (2), Kluwer Academic Publishers, Netherlands, pp. 187–197

[31] G. Welch and G. Bishop (2002). An Introduction to the Kalman Filter. *UNC-Chapel Hill, TR 95-041*.

[32] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp (2002). A tutorial on particle filters for on-line non-linear/non-Gaussian Bayesian tracking. *IEEE Trans. on Signal Process*., 50, (2), pp. 174–188

[33] A. Senior, A. Hampapur, Y.-L. Tian, L. Brown, S. Pankanti, and R. Bolle. Appearance models for occlusion handling. In *2nd IEEE Int. Workshop on PETS, Kauai, Hawaii, USA*, Dec 2001.

[34] B. Bhanu, I. Pavlidis, R. Hummel. *Guest Editorial:* Special issue on computer vision beyond the visible spectrum. Machine Vision and Applications (2000) 11: 265–266

[35] S.-S. Lin. Review: Extending visible band computer vision techniques to infrared band images. *Technical report, GRASP Laboratory, Computer and Information Science Department*, University of Pennsylvania, 2001.

[36] C. K. Eveland, D. A. Socolinsky, L. B. Wolff. Tracking human faces in infrared video. *Image and VisionComputing*, Vol. 21, pp. 579-590, 2003.

[37] M. Bertozzi, A. Broggi, P. Grisleri, T. Graf, M. Meinecke. Pedestrian Detection in Infrared Images. *Proc.IEEE Intelligent Vehicles Symposium 2003*, pp. 662-667, Columbus (USA), June 2003.

[38] F. Xu, X. Liu, and K. Fujimura. Pedestrian Detection and Tracking With Night Vision. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 6, No. 1, pp. 63-71, March 2005.

[39] J. Davis and V. Sharma. Robust detection of people in thermal imagery. In *Proc. Int. Conf. Pat. Rec.*, pages 713–716, 2004

[40] J. Han and B. Bhanu. Human activity recognition in thermal infrared imagery. *Computer Vision and Pattern Recognition*, 20-26 June, 2005.

[41] McDaniel, R., Scribner, D., Krebs, W., Warren, P., Ockman, N., McCarley, J. (1998). Image fusion for tactical applications. Proceedings of the SPIE - Infrared Technology and Applications XXIV, 3436, 685-695.

[42] J. Li. Spatial quality evaluation of fusion of different resolution images. *International Archives of Photogrammetry and Remote Sensing*, Vol.33, 2000.

[43] A. Toet. Heirarchical image fusion. *Machine Vision and Applications*, 3:1–11, 1990

[44] M. Pavel, J. Larimer, and A. Ahumada. Sensor fusion for synthetic vision. In *Conference on Computing in Aerospace, AIAA, 1991.*

[45] H. Li, B. Manjunath, and S. Mitra. Multisensor image fusion using the wavelet transform. In *Graphical Models and Image Processing*, volume 57, pages 234–245, 1995

[46] A. Utsumi, H. Mori, J. Ohya, and M. Yachida. Multipleview-based tracking of multiple humans. In *Proceedings of the 14th ICPR*, pages 597–601, 1998

[47] A. Nakazawa, H. Kato, and S.Inokuchi. Human tracking using distributed vision systems. In *Proceedings of the 14th ICPR*, pages 593–596, 1998

[48] H. Torresan, B. Turgeon, C. Ibarra-Castanedo, P. Hébert, X. Maldague,. Advanced Surveillance Systems: Combining Video and Thermal Imagery for Pedestrian Detection. In *Proc. of SPIE, Thermosense XXVI*, volume 5405 of *SPIE*, pages 506–515, April 2004.

[49] J. W. Davis and V. Sharma**.** Fusion-Based Background-Subtraction using Contour Saliency. *Computer Vision and Pattern Recognition*, 20-26 June, 2005

[50] C.O. Conaire, N. O.Connor, E. Cooke, A. Smeaton. Multispectral Object Segmentation and Retrieval in Surveillance Video. *To appear in International Conference on Image Processing,* 2006.

[51] P. Kumar, A. Mittal and P. Kumar. Fusion of Thermal Infrared and Visible Spectrum Video for Robust Surveillance. In *5th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, LNCS 4338, pp. 528 – 539, 2006.

[52] Ó. Conaire, C. O'Connor, N.E. Cooke and E. Smeaton. Comparison of fusion methods for thermo-visual surveillance tracking. *In International Conference on Information Fusion,2006.*

[53] D. L. Hall and J. Llinas. An introduction to multisensor fusion. *In Proceedings of the IEEE: Special Issues on Data Fusion,* pages 85(1):6–23, January 1997.

[54] R. R. Brooks and S.S. Iyengar. Multi-sensor fusion: fundamentals and applications with software. Upper Saddle River, N.J. : Prentice Hall PTR, 1998.

[55] S. Nazarko. Evaluation of data fusion methods using kalman filtering and TBM. Maters thesis. University of Jyväskylä. 2002.

[56] Ph. Smets. The Transferable Belief Model for Quantified Belief Representation. *In Handbook of defeasible reasoning and uncertainty management* systems. Gabbay D. M. and Smets Ph. Eds. Vol. 1, Kluwer, Doordrecht, 1998, pg. 267-30*1*.

[57] G. Loy, L. Fletcher, N. Apostolo, and A. Zelinsky. An adaptive fusion architecture for target tracking. In *IEEE International Conference on Automatic Face and Gesture Recognition (FGR)*, 2002.

[58] A. Ayoun, and P. Smets. Data association in multi-target detection using the transferable belief model. *International Journal of Intelligent Systems*, 16:1167-1182. 2001

[59] G. Powell, D. Marshall, R. Milliken, K. Markham. A Data Fusion System for Object Recognition based on Transferable Belief Models and Kalman Filters. *In `Proceedings of 7th International Conference on Information Fusion.* Sweden, Pp 54-61. 2004

[60] P.J. Escamilla-Ambrosio, N. Mort. A Hybrid Kalman Filter - Fuzzy Logic Architecture for Multisensor Data Fusion. *Proceedings of the 2001 IEEE International Symposium on Intelligent Control* , pp. 364-369 , 2001

[61] Z. Elouedi, K. Mellouli, P. Smets. Assessing sensor reliability for multisensor data fusion within the transferable belief model. *IEEE Transactions on Systems, Man and Cybernetics, Part B,* page 782-787. 2004

[62] L. Snidaro, G.L. Foresti, R. Niu, and P.K. Varshney. Sensor fusion for video surveillance. *Proceedings of the seventh international conference on information fusion,* Vol. 2, Stockholm, Sweden, June 28th-July 1st, 2004, pp. 739-746.

[63] G. Rogova and V. Nimier. Reliability in information fusion: Literature. Survey. *In Proceedings of 7th International Conference on Information. Fusion,* Sweden, pp. 1158-1165. 2004

[64] P. Kumar**,** A. Mittal and P. Kumar. A Multimodal Audio, Visible and Infrared Surveillance System (MAVISS). *In Proceedings of the 3rd IEEE International Conference on Intelligent Sensing and Information Processing (ICISIP)*, pp. 151-157, 2005.

[65] J. Hershey, H. Attias, N. Jojic, and T. Krisjianson. Audio visual graphical models for speech processing. *In IEEE International Conference on Speech, Acoustics, and Signal Processing* (ICASSP04), Motreal, Canada, May 2004.

[66] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy. Dynamic bayesian networks for audio-visual speech recognition. *In EURASIP Journal on Applied Signal Processing* (JASP02), 2002.

[67] H. J. Nock, G. Iyengar, and C. Neti. Speaker localization using audio-visual synchrony: An

empirical study. *In International Conference on Image and Video Retrieval* (CIVR03), 2003. pages 488-499.

[68] D. Gatica-Perez, G. Lathoud, I. McCowan, J. Odobez, and D. Moore. Audio-visual speaker tracking with importance particle filter. *In IEEE International Conference on Image Processing* (ICIP03), 2003.

[69] N. Checka, K. W. Wilson, M. R. Siracusa, and T. Darrell. Multiple person and speaker activity tracking with a particle filter. *In International Conference on Acoustics Speech and Signal Processing* (ICASSP04), 2004.

[70] H. J. Nock, G. Iyengar, and C. Neti. Assessing face and speech consistency for monologue detection in video. *In ACM Multimedia*, 2002.

[71] P. K. Atrey, M. S. Kankanhalli and R. Jain. Information assimilation framework for event detection in multimedia surveillance systems. *Special Issue on "Multimedia Surveillance Systems" in Springer/ACM Multimedia Systems Journal*, September 2006

[72] D. Whitney, and J. Pisano, TASC, Inc., Reading, Massachusetts. *AutoAlert: Automated Acoustic Detection of Incidents.* December 26, 1995.

[73] M. Cowling, R. Sitte. Comparison of Techniques for Environmental Sound Recognition. *Pattern Recognition Letters,* Elsevier Science Inc., Vol. 24, Issues 15, pp. 2895-2907, Nov. 2003.

[74] Proximity Sensors. http://www.machinedesign.com

[75] Z. Zhu, W. Li and G. Wolberg. Integrating LDV Audio and IR Video for Remote Multimodal Surveillance. *IEEE Workshop on Object Tracking and Classification In and Beyond the Visible Spectrum*, in conjunction with IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, June 2005

[76] I. Ulrich and J. Borenstein, "The guidecane, applying mobile robot technologies to assist the visually impaired," IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, vol. 31, no. 2, pp. 131–136, 2001.

[77] S. Shoval, I. Ulrich, and J. Borenstein, Computerized Obstacle Avoidance Systems for the Blind and Visually Impaired. Intelligent Systems and Technologies in Rehabilitation Engineering, CRC Press, 2000.

[78] Caimi, F.M. Bailey, B.C. Blatt, J.H. Undersea object detection and recognition: the use of spatially andtemporally varying coherent illumination. OCEANS '99 MTS/IEEE. Riding the Crest into the 21st Century ,pages 1474-1479 vol.3

[79] Heckman, P., and Hodgson, "Underwater Optical Range Gating", IEEE Journal of Quantum Electronics, QE-3, 11, Nov. I967

[80] A. F. Bobick, and J. W. Davis. The Recognition of Human Movement Using Temporal Templates. *PAMI*, Vol. 23, No. 3, 2001.

[81] K. Takahashi, S. Seki, H.Kojima, and R. Oka, "Recognition of dexterous manipulations from time varying images," in *Proc. IEEEWorkshop Motion of Non-Rigid and Articulated Objects*, Austin, TX, 1994, pp. 23–28.

[82] A. F. Bobick and A. D.Wilson. A state-based technique to the representation and recognition of gesture. *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 1325–1337, Dec. 1997.

[83] R. Duggad, U. B. Desai. A Tutorial on Hidden Markov Models. Technical Report No. SPANN-96.1. 1996. Indian Institute of Technology Bombay.

[84] L. Rabinier, A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of IEEE*. 77 (2) (1989) 257-285.

[85] M. Leo, P. Spagnolo, T. D'Orazio, A. Distante. Human Activity Recognition in Archaeological Sites by Hidden Markov Models. *PCM (2)* , 1019-1026, 2004

[86] D. Mahajan, N. Kwatra, S. Jain, P. Kalra, S. Banerjee. A Framework for Activity Recognition and Detection of Unusual Activities. 15-21, *ICVGIP*, 2004, Kolkata, India.

[87] M. Yang and N. Ahuja. Extraction and classification of visual motion pattern recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1998, pp. 892–897.

[88] U. Meier, R. Stiefelhagen, J. Yang, and A.Waibel. Toward unrestricted lip reading. *Int. J. Pattern Recognit. Artificial Intell.*, vol. 14, no. 5, pp. 571–585, Aug 2000

[89] O. Boiman and M. Irani. Detecting irregularities in images and in video. In *Proc. IEEE International Conference on Computer Vision*, pages 1985–1988, Beijing, China, Oct. 15-21 2005.

[90] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan. Semi-supervised adapted HMMs for unusual event detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, July 2005.

[91] P. Kumar, A. Mittal and P. Kumar. A Multi-modal Data Fusion Framework Using Transferable Belief Model and Kalman filter for Robust Tracking in Dynamic Environment. Communicated to *Signal, Video and Image processing Journal,* Springer Verlag publication.