16

A Survey on Behaviour Analysis in Video Surveillance Applications

Teddy Ko
Raytheon Company,
USA

1. Introduction

There is an increasing desire and need in video surveillance applications for a proposed solution to be able to analyze human behaviors and identify subjects for standoff threat analysis and determination. The main purpose of this survey is to look at current developments and capabilities of visual surveillance systems and assess the feasibility and challenges of using a visual surveillance system to automatically detect abnormal behavior, detect hostile intent, and identify human subject.

Visual (or video) surveillance devices have long been in use to gather information and to monitor people, events and activities. Visual surveillance technologies, CCD cameras, thermal cameras and night vision devices, are the three most widely used devices in the visual surveillance market. Visual surveillance in dynamic scenes, especially for humans, is currently one of the most active research topics in computer vision and artificial intelligence. It has a wide spectrum of promising public safety and security applications, including access control, crowd flux statistics and congestion analysis, human behavior detection and analysis, etc.

Visual surveillance in dynamic scene with multiple cameras, attempts to detect, recognize and track certain objects from image sequences, and more importantly to understand and describe object behaviors. The main goal of visual surveillance is to develop intelligent visual surveillance to replace the traditional passive video surveillance that is proving ineffective as the number of cameras exceed the capability of human operators to monitor them. The goal of visual surveillance is not only to put cameras in the place of human eyes, but also to accomplish the entire surveillance task as automatically as possible. The capability of being able to analyze human movements and their activities from image sequences is crucial for visual surveillance.

In general, the processing framework of an automated visual surveillance system includes the following stages: Motion/object detection, object classification, object tracking, behavior and activity analysis and understanding, person identification, and camera handoff and data fusion.

Almost every visual surveillance system starts with motion and object detection. Motion detection aims at segmenting regions corresponding to moving objects from the rest of an image. Subsequent processes such as object tracking and behavior analysis and recognition are greatly dependent on it. The process of motion/object detection usually involves background/environment modeling and motion segmentation, which intersect each other

during the processing. Motion segmentation in image sequences aims at detecting regions corresponding to moving objects such as humans or vehicles. Detecting moving regions provides a focus of attention for later processes such as tracking and behavior analysis as only these regions need be considered and further investigated.

After motion and object detection, surveillance systems generally track moving objects from one frame to another in an image sequence. The tracking algorithms usually have considerable intersection with motion detection during processing. Tracking over time typically involves matching objects in consecutive frames using features such as points, lines or blobs.

Behavior understanding involves analysis and recognition of motion patterns, and the production of high-level description of actions and interactions between or among objects. In some circumstances, it is necessary to analyze the behaviors of people and determine whether their behaviors are normal or abnormal.

The problem of who enters the area and/or engages in an abnormal or suspicious act under surveillance is of increasing importance for visual surveillance. Human face and gait are now regarded as the main biometric features that can be used for personal identification in visual surveillance systems.

Motion detection, tracking, behavior understanding, and personal identification at a distance can be realized by single camera-based visual surveillance systems. Multiple camera-based visual surveillance systems can be extremely helpful because the surveillance area is expanded and multiple view information can overcome occlusion. Tracking with a single camera easily generates ambiguity due to occlusion or depth. This ambiguity may be eliminated from another view. However, visual surveillance using multiple cameras also brings problems such as camera installation (how to cover the entire scene with the minimum number of cameras), camera calibration, object matching, automated camera switching, and data fusion.

The video process of surveillance systems has inherited same difficult challenges while approaching a computer vision application, i.e., illumination variation, viewpoint variation, scale (view distance) variation, and orientation variation. Existing surveillance solutions to object detection, tracking, and identification from video problems tend to be highly domain specific. An indication of the difficulty of creating a single general purpose surveillance system comes from the video surveillance and monitoring (VSAM) project at CMU (Collins et al., 2000) and other institutions (Borg et al., 2005; PETS, 2007). VSAM at CMU is one of the most ambitious surveillance projects yet undertaken, and has advanced the state of the art in many areas of surveillance research. This project was intended as a general purpose system for automated surveillance of people and vehicles in cluttered environments, using a range of sensors including color CCD cameras, thermal cameras, and night vision cameras. However, due to the difficulty of developing general surveillance algorithms, a visual surveillance system usually has had to be designed as a collection of separate algorithms, which are selected on a case by case basis.

The flow and organization of this review paper has followed four very thorough, excellent surveys conducted by (Ko, 2008; Wang et al., 2003; Hu et al., 2004; Kumar et al., 2008) when discussing the general framework of automated visual surveillance systems as shown in Fig. 1, enriching with the general architecture of a video understanding system (Bremond et al., 2006) in behavior analysis and with expandable network system architecture as illustrated in (Cohen et al., 2008). The main intent of this paper is to give engineers, scientists, and/or managers alike, a high-level, general understanding of both the theoretical and practical perspectives involved with a visual surveillance system and its potential challenges while considering implementing or integrating a visual surveillance system.

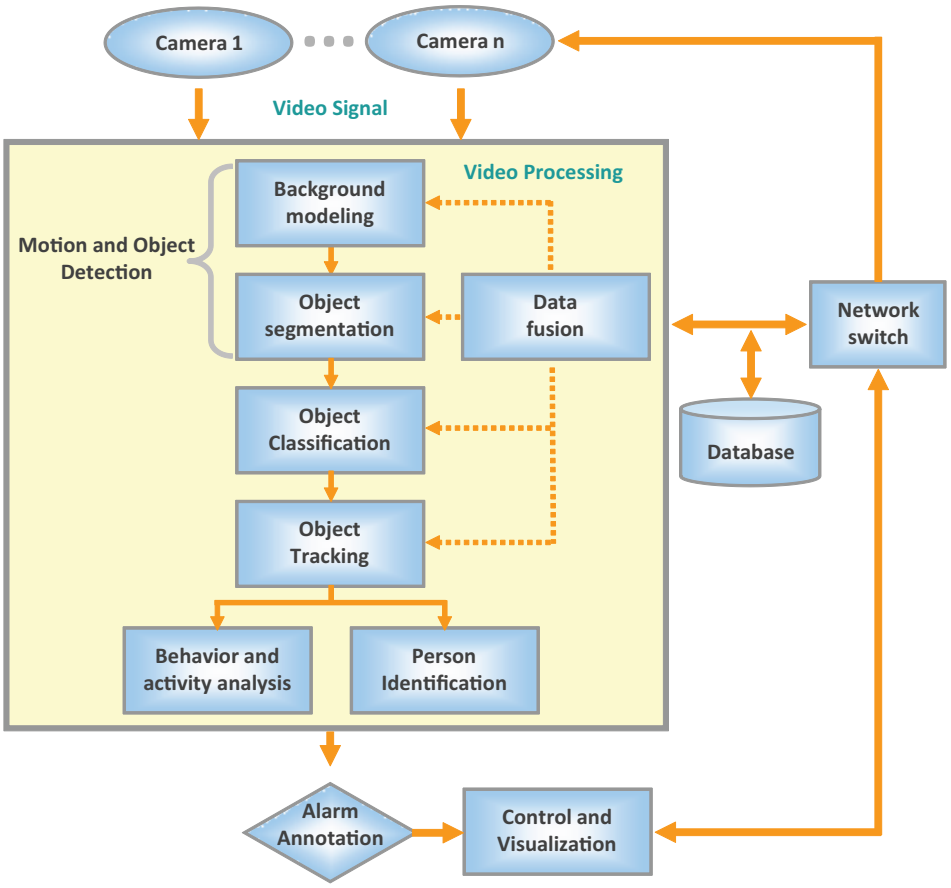


Fig. 1. A general framework of an automated visual surveillance system

This paper reviews and exploits developments and general strategies of stages involved in video surveillance, and analyzes the challenges and feasibility for combining object tracking, motion analysis, behavior analysis, and biometrics for stand-off human subject identification and behavior understanding. Behavior analysis using visual surveillance involves the most advanced and complex researches in image processing, computer vision, and artificial intelligence. There were many diverse methods (Saligrama et al., 2010) have been used while approaching this challenge; and they varied and depended on the required speed, the scope of application, and resource availability, etc. The motivation of writing and presenting a survey paper on this topic instead of a how-to paper for a domain specific application is to review and gain insight in visual surveillance systems from a big picture first. Reviewing/surveying existing available works to enable us to understand and answer the following questions better: Developments and strategies of stages involved in a general visual surveillance system; how to detect and analyze behavior and intent; and how to approach the challenge, if we have opportunities.

2. Motion and object detection

Most visual surveillance systems start with motion detection. Motion detection methods attempt to locate connected regions of pixels that represent the moving objects within the scene; different approaches include frame-to-frame difference, background subtraction and motion analysis using optical flow techniques. Motion detection aims at segmenting regions corresponding to moving objects from the rest of an image. The motion and object detection process usually involves environment (background) modeling and motion segmentation. Subsequent processes such as object classification, tracking, and behavior recognition are greatly dependent on it.

Most of segmentation methods use either temporal or spatial information in the image sequence. Several widely used approaches for motion segmentation include temporal differencing, background subtraction, and optical flow.

Temporal differencing makes use of the pixel-wise difference between two to three consecutive frames in an image sequence to extract moving regions. Temporal differencing is very fast and adaptive to dynamic environments, but generally does a poor job of extracting all the relevant pixels, e.g., there may be holes left inside moving entities.

Background subtraction is very popular for applications with relatively static backgrounds as it attempts to detect moving regions in an image by taking the difference between the current image and the reference background image in a pixel-by-pixel fashion. However, it is extremely sensitive to changes of environment lighting and extraneous events. The numerous approaches to this problem differ in the type of background model and the procedure used to update the background model. The estimated background could be simply modeled using just the previous frame; however, this would not work too well. The background model at each pixel location could be based on the pixel's recent history. Background subtraction methods store an estimate of the static scene, accumulated over a period of observation; this background model is used to find foreground (i.e., moving objects) regions that do not match the static scene. Recently, some statistical methods to extract change regions from the background are inspired by the basic background subtraction methods as described above. The statistical approaches use the characteristics of individual pixels or groups of pixels to construct more advanced background models, and the statistics of the backgrounds can be updated dynamically during processing. Each pixel in the current image can be classified into foreground or background by comparing the statistics of the current background model. This approach is becoming increasingly popular due to its robustness to noise, shadow, changing of lighting conditions, etc. (Stauffer & Grimson, 1999).

Optical flow is the velocity field, which warps one image into another (usually very similar) image, and is generally used to describe motion of point or feature between images (Watson & Ahumada, 1985). Optical flow methods are very common for assessing motion from a set of images. However, most optical flow methods are computationally complex, sensitive to noise, and would require specialized hardware for real-time applications.

3. Object classification

Different moving regions may correspond to different moving objects in natural scenes. To further track objects and analyze their behaviors, it is essential to correctly classify moving objects. For instance, the moving objects are humans, vehicles, or objects of interest of an investigated application. Object classification can be considered as a standard pattern

recognition task. There are two main categories of approaches for classifying moving objects: shape-based classification and motion-based classification
Different descriptions of shape information of motion regions such as points, boxes, silhouettes and blobs are available for classifying moving objects. In general, human motion exhibits a periodic property, so this has been used as a strong cue for classification of moving objects also.

4. Object tracking

The task of tracking objects as they move in substantial clutter, and to do it at, or close to, video frame-rate is challenging. The challenge occurs if elements in the background mimic parts of features of the foreground objects. In the most severe case, the background may consist of objects similar to the foreground object(s), e.g., when a person is moving past a person, a group of people, or a crowd (Cavallaro et al., 2005).
The object tracking module is responsible for the detection and tracking of moving objects from individual cameras; object locations are subsequently transformed into 3D world coordinates. The camera handoff and data fusion module (or algorithm) then determines single world measurements from the multiple observations. Object tracking can be described as a correspondence problem and involves finding which object in a video frame related to which object in next frame (Javed & Shah, 2002). Normally, the time interval between two successive frames is small, thus the inter-frame changes are limited, allowing the use of temporal constraints and/or object features to simplify the correspondence problem. Tacking methods can be roughly divided into four major categories, and algorithms from different categories can be integrated together (Cavallaro et al., 2005, Javed & Shah, 2002).

a. Region-based Tracking

Region-based tracking algorithms track objects according to variation of the image regions corresponding to the moving objects. For these algorithms, the motion regions are usually detected by subtracting the background from the current images.

b. Contour-based Tracking

In contour-based methods instead of tracking the whole set of pixels comprising an object, the algorithms track only the contour of the object (Isard & Blake, 1996).

c. Feature-based Tracking

Feature-based methods use features of a video subject to track parts of the object. Feature-based tracking algorithms perform recognition and tracking of objects by extracting elements, clustering them into higher level features and then matching the feature between images.

d. Model-based Tracking

Model-based tracking algorithms track objects by matching projected object model. The models are usually constructed off-line with manual measurement, CAD tools or computer vision techniques. Generally, model-based human body tracking involves three main tasks: 1) construction of human body models; 2) representation of a priori knowledge of motion models and motion constraints; and 3) prediction and search strategies. Construction of human body models is the base of model-based human tracking. In general, the more complex a human body model, the more accurate the tracking results, but the more expensive the computation. Traditionally, the geometry structure of a human body can be represented in four styles: Stick figure, 2-D contour, volumetric model, and hierarchical model.

e. Hybrid Tracking

Hybrid approaches are designed as a hybrid between region-based and feature-based techniques. They exploit the advantages of two by considering first the object as an entity and then by tracking its parts.

5. Extraction and motion information

Before discussing the details of the extraction of motion information, Fig. 3 shows how a surveillance system may extract and learn motion patterns, e.g., a walk cycle, using an example of 4-level decomposition of the human dynamics as illustrated in (Bregler, 1997). Each level represents a set of random variables and probability distributions over hypotheses. The lowest level is a sequence of input images. For each pixel, we represent the spatio-temporal image gradient and optionally the color value as a random variable. The second level shows the blob hypotheses. Each blob is represented with a probability distribution over coherent motion (rotation and translation or full affine motion), color (HSV values), and spatial "support-regions". In the third level, temporal sequences of blob tracks are grouped to linear stochastic dynamical models. At the fourth and highest level, each dynamic model corresponds to the emission probability of the state of a Hidden Markov Model (HMM).

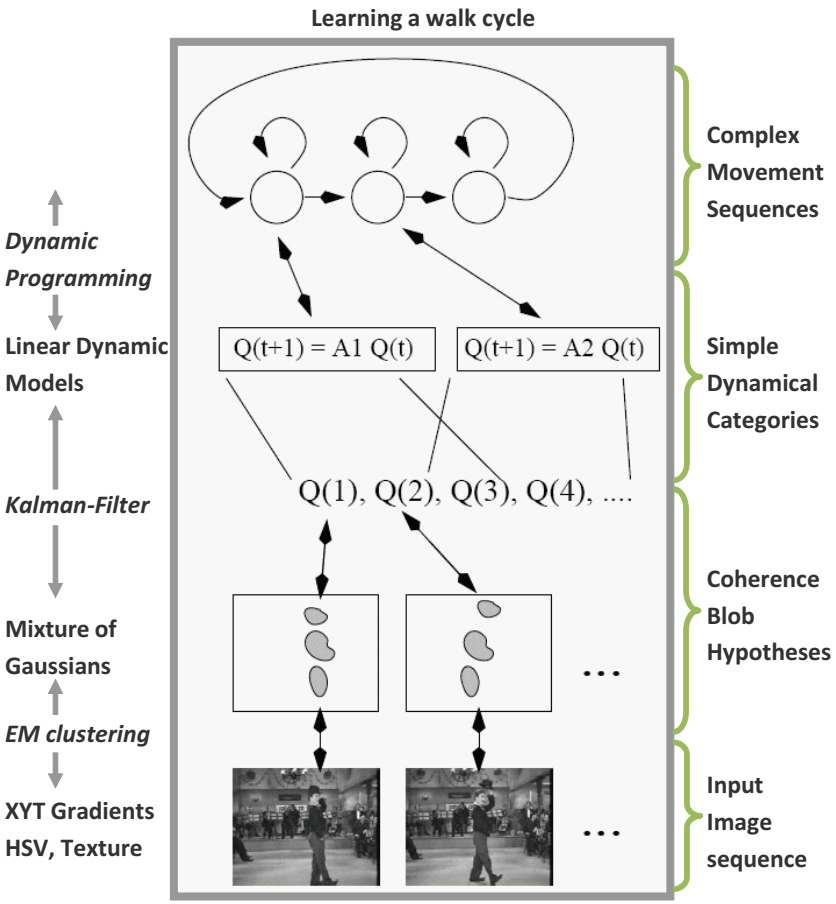


Fig. 2. Learning and recognizing human dynamics in video sequences (Bregler, 1997)

For example, the movement of one leg during a walk cycle can be decomposed into one coherent motion blob for the upper leg, and one coherent motion blob for the lower leg; one dynamic system for all the time frames while the leg has ground support, and one dynamic system in case the leg is swinging above ground, and a "cycle" HMM with multiple states. The state space of the dynamic systems is the translation and angular velocities of the blob hypothesis. The HMM stays in the first state for as many frames as the first dynamical system is valid, transitions to the second state once the second dynamic system is valid, and then cycles back to the first state for the next walk cycle.

The first important step in motion-based recognition is the extraction of motion information from a sequence of images. Motion perception and interpretation plays a very important role in a visual surveillance system. There are generally three methods for extracting motion information from a sequence of images: Optical flow, trajectory-based features, and region-based features.

a. Optical Flow Features

Optical flow methods are very common for assessing motion from a set of images. Optical flow is an approximation of the two-dimensional flow field from image intensities. Optical flow is the velocity field, which warps one image into another (usually very similar) image. Several methods have been developed, however, accurate and dense measurements are difficult to achieve (Cedras & Shah, 1995).

b. Trajectory-based Features

Trajectories, derived from the locations of particular points on an object in time, are very popular because they are relatively simple to extract and their interpretation is obvious (Morris & Trivedi, 2008). The generation of motion trajectories from a sequence of images typically involves the detection of tokens in each frame and the correspondence of such tokens from one frame to another. The tokens need to be distinctive enough for easy detection and stable through time so that they can be tracked. Tokens include edges, corners, interest points, regions, and limbs. Several proposed solutions (Cavallaro et al., 2005; Koller-meier & Van Gool, 2001; Makris & Ellis, 2005; Bobick & Wilson, 1997) for human actions modeling and recognition using the trajectory-based features approach. In the first step, an arbitrary changing number of objects are tracked. From the history of the tracked object states, temporal trajectories are formed which describe the motion paths of these objects. Secondly, characteristic motion patterns are learned by e.g. clustering these trajectories into prototype curves. In the final step, motion recognition is then tackled by tracking the position within these prototype curves based on the same method used for the object tracking.

c. Region- or Image-based Features

For certain types of objects or motions, the extraction of precise motion information for each single point is neither desirable nor necessary. Instead, the ability to have a more general idea about the content of a frame might be sufficient. Features generated from the use of information over a relatively large region or over the whole image are referenced here as region-based features. This approach has been used in several studies (Jan, 2004).

6. Behaviour analysis and understanding

One of most difficult challenges in the domain of computer vision and artificial intelligence is semantic behavior learning and understanding from observing activities in video (visual) surveillance. The research in this area concentrates mainly on the development of methods

for analysis of visual data in order to extract and process information about the behavior of physical objects (e.g., humans) in a scene.

In automated visual surveillance systems, reliable detection of suspicious or endangering human behavior is of great practical importance (Regazzoni et al., 2010; Lao et al., 2010). An automated visual surveillance system generally requires a reliable combination of image processing and artificial intelligence techniques. Image processing techniques are used to provide low level image features. Artificial intelligence techniques are used to provide expert decisions. Extensive research has been reported on low level image processing techniques such as object detection, recognition, and tracking; however, relatively few researches has been reported on reliable classification and understanding of human activities from the video image sequences.

Detection of suspicious human behavior involves modeling and classification of human activities with certain rules. Modeling and classification of human activities are not trivial due to the randomness and complex nature of human movement. The idea is to partition the observed human movements into some discrete states and then classify them appropriately. Apparently, partitioning of the observed movements is very application-specific and overall hard to predict what will constitute suspicious or endangering behavior (Cohen et al., 2008; Jan, 2004; Saligrama et al., 2010).

Most approaches in the field of video understanding incorporated methods for detection of domain-specific events (Bremond et al., 2006). Examples of such systems use dynamic time warping for gesture recognition (Bobick & Wilson, 1997) or self-organizing networks for trajectory classification (Ivanov & Bobick, 2000; Bobick & Davis, 2001). The main drawback of these approaches is the usage of techniques that are specific only for a certain application domain which causes difficulties when applying these techniques to other areas (Bremond et al., 2006). Therefore, some researchers (Bremond et al., 2006; Ivanon & Bobick, 2000) have proposed and adopted a two-step approach to the problem of video understanding:

- A lower-level image processing visual module is used to extract visual cues and primitive events
- This collected information is used in a higher-level artificial intelligence module for the detection of more complex and abstract behavior patterns

By dividing the problem into two or three sub-problems, researchers can use simpler and more domain-independent techniques in each stage. The first stage usually involves and uses image processing and stochastic techniques for data analysis while the second stage conducts structural analysis of the symbolic data gathered at the previous step.

In the general visual surveillance process framework as shown in Fig. 1, the motion detection/segmentation and object classification are usually grouped as lower-level vision tasks. Human behavior recognition is based on successfully tracking the human subject through image sequences, and is considered a high-level vision task. The tracking process as discussed in (Wang et al., 2003) can be considered an intermediate-level vision task, or it can be split into lower and higher two stages as proposed in (Bremond et al., 2006) and shown in Fig. 3.

As shown in Fig. 3, at the first level of a general video surveillance system, geometric features, like areas of motions, are extracted. Based on those extractions, objects are recognized and tracked. At the second level, events in which the detected objects participate are recognized. For performing this task, a selected representation of events is used that defines concepts and relations in the domain of human activity monitoring.

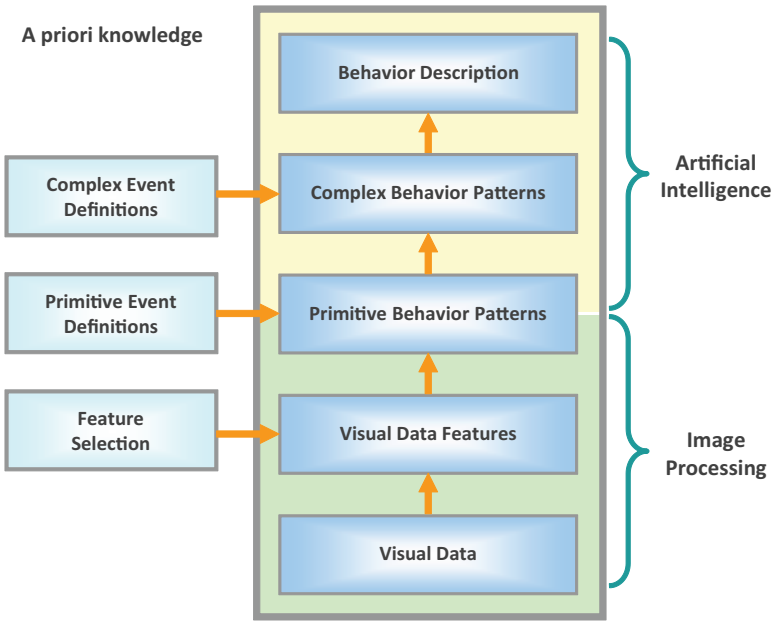


Fig. 3. A general architecture of a video understanding system.

For the computer vision community, a natural approach to recognize scenarios consists of using a probabilistic or neural network. The nodes of this network correspond usually to scenarios that are recognized at a given instance with a computed likelihood.

For the artificial intelligence community, a natural way to recognize a scenario is to use a symbolic network where nodes correspond usually to the Boolean recognition of scenarios. The common characteristic of these approaches is that all totally recognized behaviors are stored.

Another development that has captured the attention of researchers, is the unsupervised behavior learning and recognition, consisting of the capability of a vision interpretation system of learning and detecting the frequent scenarios of a scene without requiring the prior definitions of behaviors by the user.

Any scene object involved in a behavior/action should also include other individuals, groups of people, crowds, or static objects (e.g., equipments). Activities involve a regularly repeating sequence of motion events. The automatic video understanding and interpretation needs to know how to represent and recognize behaviors corresponding to different types of concepts, which include (Bremond et al., 2006; Medioni et al., 2001; Levchuk et al., 2010):

- **Basic Properties**: A basic property is a characteristic of an object such as its trajectory or speed.
- **States**: A state describes a situation characterizing one or several objects (actors) defined at given time (e.g., a subject is agitated) or a stable situation defined over a time interval. For the state: "an individual stays close to the ticket vending machine," two subjects (actors) are involved: an individual and a piece of equipment.
- **Events**: An event is a change of state at two consecutive times (e.g., a subject enters an area of interest).

- **Scenarios**: A scenario is a combination of states, events or sub-scenarios. Behaviors are specific scenarios, dependent on the application defined by the users. For example, to monitor metro stations, end-users could have defined targeted behaviors: "Loitering", "Unattended Luggage", "Vandalism", "Overcrowding", "Fighting", etc.

The ability to extract semantic information from human biologic motion has been known for some time. In his seminal work, Johansson (1973) revealed that presenting coordinated human joint motion was sufficient for rendering the impression of a human being walking or running through space.

With respect to detecting hostile intent (Cohen et al., 2008), each point in the point-light walker (PLW) might have its own gesture motion – which when examined in relation to the links in the object, can be used to determine the overall state of the system. Unusual events such as vandalism or overcrowded areas can be detected by unusual movements as well as unlikely object positions.

People have had the innate ability to recognize others' emotional dispositions based on intuition; this innateness must also manifest itself physically. For instance, when someone is experiencing emotion, what visual cues exist that communicate this? Facial expression, is an immediate indicator, but what about their behavior? Does posture, gesture, or specific body parts communicate this also? A system will be able to learn the visual cues found to be of some significance in identifying an emotion (Johansson, 1973) by identifying specific regions of the body that identify emotions,. Researchers will discover that motions of certain body parts may identify an emotion more than others (Cohen et al., 2008; Johansson, 1973; Montepare et al., 1987). For instance, researchers may discover that in anger the torso is most evocative of that emotion.

The review of available and state of the art techniques show the large diversity of video understanding techniques in automatic behavior recognition. The challenge is to efficiently combine these techniques to address the large diversity of the real world. Behavior pattern learning and understanding may be thought of as the classification of time varying feature data, i.e., matching an unknown test sequence with a group of labeled reference sequences representing typical or learned behaviors (Bobick & Davis, 2001). The fundamental problem of behavior understanding is to learn the reference behavior sequences from training samples, and to devise both training and matching methods for coping effectively with small variations of the feature data within each class of motion pattern. The major existing methods for behavior understanding include the following:

a. Hidden Markov Models (HMMs): A HMM is a statistical tool used for modeling generative sequences characterized by a set of observable sequences (Brand & Kettnaker, 2000).
b. Dynamic Time Warping (DTM): DTW is a technique that computes the non-linear warping function that optimally aligns two variable length time sequences (Bobick & Wilson, 1997). The warping function can be used to compute the similarity between two time series or to find corresponding regions between the two time series.
c. Finite-State Machine (FSM): FSM or finite-state automaton or simply a state machine, is a model of behavior composed of a finite number of states, transitions between those states, and actions. A finite state machine is an abstract model of a machine with a primitive internal memory.
d. Nondeterministic-Finite-State Automaton (NFA): A NFA or nondeterministic finite state machine is a finite state machine where for each pair of state and input symbols,

there may be several possible next states. This distinguishes it from the deterministic finite automaton (DFA), where the next possible state is uniquely determined. Although the DFA and NFA have distinct definitions, it may be shown in the formal theory that they are equivalent, in that, for any given NFA, one may construct an equivalent DFA, and vice-versa.

e. Time-Delay Neural Network (TDNN): TDNN is an approach to analyzing time-varying data. In TDNN, the 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 methods have been successfully applied to applications, such as hand gesture recognition and lip reading.

f. Syntactic/Grammatical Techniques: The basic idea in this approach is to divide the recognition problem into two levels. The lower level is performed using standard independent probabilistic temporal behavior detectors, such as HMMs, to output possible low-level temporal features. These outputs provide the input stream for a stochastic context-free grammar parser. The grammar and parser provide longer range temporal constraints, disambiguate uncertain low-level detection, and allow the inclusion of a priori knowledge about the structure of temporal behavior (Ivanov & Bobick, 2000).

g. Self-Organizing Neural Network: The methods discussed in (a) - (f) all involve supervised learning. They are applicable for known scenes where the types of object motions are already known. The self-organizing neural networks are suited to behavior understanding when the object motions are unrestricted.

h. Agent-Based Techniques: Instead of learning large amounts of behavior patterns using a centralized approach, agent-based methods decompose the learning into interactions of agents with much simpler behaviors and rules (Bryll et al., 2005).

i. Artificial Immune Systems: Several researchers have exploited the feasibility of learning behavior patterns and hostile intents in the optical flow level using artificial immune system approaches (Sarafijanovic & Leboudec, 2004).

7. Person identification

In most of video surveillance system literatures, the person identification is achieved by motion analysis and matching, such as gait, gesture, posture analysis and comparison (Hu et al., 2004). In model-based methods, parameters for gait, gesture, and/or posture, such as joint trajectories, limb lengths, and angular speeds are measured. Statistical recognition techniques usually characterize the statistical description of motion image sets and have been well developed in automatic gait recognition. Physical-parameter-based methods make use of geometric structural properties of a human body to characterize a person's gait pattern. The parameters used included height, weight, stride cadence, length, etc. For motion recognition based on spatio-temporal analysis, the action or motion is characterized via the entire 3-D spatio-temporal data volume spanned by the moving person in the image sequence.

Human gait and face are now regarded as the main biometric features that can be used for personal identification in visual surveillance systems. The fusion of gait and face information with other standoff biometrics to further increase recognition robustness and reliability has been exploited by new surveillance systems. The problem of who is (are) now

in the area, is (are) engaging in an abnormal/suspicious act under surveillance is of increasing important for visual surveillance.

8. Camera handoff and data fusion

To expand the surveillance area and provide multiple view information to overcome, most of visual (or video) surveillance systems are multiple camera-based. In a multi-camera surveillance system, with overlapping fields of view to track objects and recognize their activities predefined by a set of activities or scenarios, or even learns new behavior patterns or new knowledge. Each camera agent performs per frame detection and tracking of scene objects, and the output data is transmitted to a centralized server where data associated and fused object tracking is performed. This tracking result is fed to a video event recognition module where spatial and temporal events relating to the objects are detected and analyzed. Tracking with a single camera easily generates ambiguity due to occlusion or depth. This ambiguity may be eliminated from another view. However, visual surveillance using multiple cameras also brings problems such as camera installation (how to cover the entire scene with the minimum number of cameras), camera calibration, object matching, automated camera switching, and data fusion (Collins et al., 2000).

Most of proposed systems use cameras as the sensor since the camera can provide resolution needed for accurate classification and position measurement. The disadvantage of image-only detection systems is the high computational cost associated with classifying a large number of candidate image regions. Accordingly, it has been a trend for several years to use a hierarchical detection structure combining different sensors. In the first step low computational cost sensors identify a small number of candidate regions of interest (ROI). LIDAR (**Li**ght **D**etection **a**nd **R**anging) is an optical remote sensing technology that measures properties of scattered light to find range and/or other information of a distant target. The prevalent method to determine distance to an object or surface is to use laser pulses. Like the similar RADAR technology, which uses radio waves instead of light, the range to an object is determined by measuring the time delay between transmission of a pulse and detection of the reflected signal. As shown in (Szarvas et al., 2006; Premebida et al., 2007), the region of interest (ROI) detector in their proposed systems receives the signal from the LIDAR sensor and outputs a list of boxes in 3 dimensional (3D) world-coordinates. The 3D ROI-boxes are obtained by clustering the LIDAR measurements. Each 3D box is projected to the image plane using the intrinsic and extrinsic camera parameters.

9. Performance evaluation

The methods of evaluating the performance of object detection, object tracking, object classification, and behavior and intent detection and identification in a visual surveillance system are more complex than some of the well-established biometrics identification applications, such as fingerprint or face, due to unconstrained environments and the complexity of challenge itself. Performance Evaluation for Tracking and Surveillance (PETS) is a good starting place when looking into performance evaluation (PETS, 2007). As shown in Fig. 4, PETS has several good data sets for both indoor and outdoor tracking evaluation and event/behavior detection.

Fig. 4. Surveillance scenario dataset shows sample images captured from multiple cameras.

PETS datasets, starting from 2000 to 2007, include:

- Outdoor people and vehicle tracking using single or multiple cameras,
- Indoor people tracking (and counting) and hand posture classification,
- Annotation of a smart meeting, including facial expression, gaze and gesture/action,
- Multiple sensor (camera) sequences for unattended luggage,
- Multiple sensor (camera) sequences for attended luggage removal (theft), and
- Multiple sensor (camera) sequences for loitering.

In addition to surveillance datasets, there are efforts, like TRECVid Evaluation (Smeaton et al., 2009), with the goal to support the development of technologies to detect visual events through standard test datasets and evaluation protocols.

10. Conclusions

Visual (or video) surveillance systems have been around for a couple of decades. Most current automated video surveillance systems can process video sequence and perform almost all key low-level functions, such as motion detection and segmentation, object tracking, and object classification with good accuracy. Recently, technical interest in video surveillance has moved from such low-level functions to more complex scene analysis to detect human and/or other object behaviors, i.e., patterns of activities or events, for standoff threat detection and prevention.

Existing behavior/event analysis systems focus on the predefined events/behaviors, e.g., to combine the results of an automated video surveillance system with spatiotemporal reasoning about each object relative to the key background regions and other objects in the scene. Advanced behavior/event analysis systems have begun to exploit the capability to automatically capture and define (learn) new behaviors/events by pattern discovery, and

further present the behavior/events to the specialists for confirmation. The increasing need for sophisticated video surveillance systems and the move to digital video surveillance infrastructure, has transformed automated video surveillance into a large scale data analysis and management challenge (Brown et al., 2006).

This paper reviews and exploits developments and general strategies of stages involved in video surveillance and analyzes the challenges and feasibility for combining object tracking, motion analysis, behavior analysis, and biometrics for stand-off human subject identification and behavior understanding. Behavior analysis using visual surveillance involves the most advanced and complex researches in image processing, computer vision, and artificial intelligence. There were many diverse methods have been used while approaching this challenge; and they varied and depended on the required speed, the scope of application, and resource availability, etc. The motivation of writing and presenting a survey paper on this topic instead of a how-to paper for a domain specific application is to review and gain insight in visual surveillance systems from a big picture first. Reviewing/surveying existing available works to enable us to understand and answer the following questions better: Developments and strategies of stages involved in a general visual surveillance system; how to detect and analyze behavior and intent; and how to approach the challenge, if we have opportunities.

11. References

Bobick, A. & Wilson, A. (1997). "A State-Based Approach to the Representation and Recognition of Gesture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 12, pp. 1325-1337.

Bobick, A. & Davis, J. (2001). "The Recognition of Human Movement Using Temporal Templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 3, pp. 257-267.

Borg, M., Thirde, D., Ferryman, J., Fusier, F., Valentin, V., Bremond, F. & Thonnat, M. (2005). "Video Surveillance for Aircraft Activity Monitoring," IEEE Conference on Advanced Video and Signal Based Surveillance, pp. 16-21.

Brand, M. & Kettnaker, V. (2000). "Discovery and Segmentation of Activities in Video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, pp. 844-851.

Bregler, C. (1997). "Learning and Recognizing Human Dynamics in Video Sequences," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, pp. 568-574.

Brown, L., Hampapur, A., Connell, J., Lu, M., Senior, A., Shu, C. & Tian, Y. (2005). "IBM Smart Surveillance System (S3): An open and extensible architecture for smart video surveillance."

Bremond, F., Thonnat, M. & Zuniga, M. (2006). "Video-understanding framework for automatic behavior recognition," *Behavior Research Methods*, Vol. 30, No. 3, pp. 416-426.

Bryll, R., Rose, R. & Quek, F. (2005). "Agent-Based Gesture Tracking," *IEEE Transactions on Systems, Man and Cybernetics*, Part A, Vol. 35, No. 6, pp. 795-810.

Cavallaro, A., Steiger, O. & Ebrahimi, T. (2005). "Tracking video objects in cluttered background," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 15, No. 4, pp. 575-584.

Cedras, C. & Shah, M. (1995). "Motion-Based Recognition: A Survey," *Image and Vision Computing*, Vol. 13, No. 2, pp. 129-155.

Cohen, C., Morelli, F. & Scott, K. (2008). "A Surveillance System for Recognition of Intent within Individuals and Crowds," *IEEE Conference on Technologies for Homeland Security*, Waltham, MA, pp. 559-565.

Collins, R., Lipton, A., Kanade, T., Fujiyoshi, H., Duggins, D., Yin, Y., Tolliver, D., Enomoto, N. & Hasegawa, O. (2000). "A System for Video Surveillance and Monitoring," Technical Report CMU-RI-TR-00-12, Carnegie Mellon University.

Dick, A. & Brooks, M. (2003). "Issues in Automated Visual Surveillance," in *Proceedings of International Conference on Digital Image Computing: Techniques and Application*, pp. 195-204.

Hu, W., Tan, T., Wang, L. & Maybank, S. (2004). "A Survey on Visual Surveillance of Object Motion and Behaviors," *IEEE Transactions on Systems, Man, and Cybernetics*, Part C: Application and Review, Vol. 34, No. 3, pp. 334-352.

Isard, M. & Blake, A. (1996). "Contour tracking by stochastic propagation of conditional density," in *Proceedings of European Conference on Computer Vision*, Cambridge, UK, pp. 343-356.

Ivanov, Y. & Bobick, A. (2000). "Recognition of Visual Activities and Interactions by Stochastic Parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, pp. 852-872.

Jan, T. (2004). "Neural Network Based Threat Assessment for Automated Visual Surveillance," in *Proceedings of IEEE International Joint Conference on Neural Networks*, Vol. 2, pp. 1309-1312.

Javed, O. & Shah, M. (2002). "Tracking and Object Classification for Automated Surveillance," *Proceedings of the 7th European Conference on Computer Vision*, Part-IV, pp. 343-357.

Johansson, G. (1973). "Visual perception of biological motion and a model for its analysis," *Perception and Psychophysics*, Vol. 14, No. 2, pp. 201-211.

Ko, T. (2008). "A Survey on behavior analysis in video surveillance for homeland security application," *AIPR*, pp. 1-8, 37th IEEE Applied Imagery Pattern Recognition Workshop.

Koller-meier, E. & Van Gool, L. (2001). "Modeling and recognition of human actions using a stochastic approach," in *Proceedings of 2nd European Workshop on Advanced Video-Based Surveillance Systems*, London, UK, pp. 17-28.

Kosmopoulos, D. & Chatzis, S. (2010). "Robust Visual Behavior Recognition: A framework based on holistic representations and multicamera information fusion," *IEEE Signal Processing Magazine*, Vol. 27, No. 5, pp. 34-45.

Kumar, P., Mittal, A. & Kumar, P. (2008). "Study of Robust and Intelligent Surveillance in Visible and Multi-modal Framework," *Informatica* 32, pp. 63-77.

Lao, W., Han, J. & With, P. (2010). "Flexible Human Behavior Analysis Framework for Video Surveillance Application," *International Journal of Digital Multimedia Broadcasting*, Vol. 2010, Article ID 920121, 9 pages.

Levchuk, G., Bobick, A. & Jones, E. (2010). "Activity and function recognition for moving and static objects in urban environments from wide-area persistent surveillance inputs," *Proc. SPIE 7704*, p. 77040P.

Makris, D. & Ellis, T. (2005). "Learning Semantic Scene Models From Observing Activity in Visual Surveillance," *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, Vol. 35, No. 3, pp. 397- 408.

Medioni, G.. Cohen, I, Bremond, F., Hongeng, S. & Nevatia, R. (2001). "Event Detection and Analysis from Video Streams," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 8, pp. 873-889.

Montepare, J., Goldstein, S. & Clausen, A. (1987). "The Identification of Emotions from Gait Information," *Journal of Nonverbal Behavior*, 11(1), pp. 33-42.

Morris, B. & Trivedi, M. (2008). "A Survey of Vision-Based Trajectory Learning and Analysis for Surveillance," *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 18, No. 8, pp. 1114-1127.

PETS. (2007). Performance Evaluation and Tracking and Surveillance (PETS) 2007; Web site: http://pets2007.net/

Premebida, C., Monteiro, C., Nunes, U. & Peixoto, P. (2007). "A Lidar and Vision-based Approach for Pedestrian and Vehicle Detection and Tracking," *IEEE Intelligent Transportation Systems Conference*, pp. 1044-1049.

Regazzoni, C., Cavallaro, A., Wu, Y., Konrad, J. & Hampapur, A. (2010). "Video Analytics for Surveillance: Theory and Practice," *IEEE Signal Processing Magazine*, Vol. 27, No. 5, pp. 16-17.

Saligrama, V., Konrad, J. & Jodoin, P.-M. (2010). "Video Anomaly Identification: A Statistical Approach," *IEEE Signal Processing Magazine*, Vol. 27, No. 5, pp. 18-33.

Sarafijanovic, S. & Leboudec, J.-Y. (2004). "An Artificial Immune System for Misbehavior Detection in Mobile Ad-Hoc Networks with Virtual Thymus, Clustering, Danger Signal and Memory Detectors," in *Proceedings of ICARIS-2004* (Third International Conference on Artificial Immune Systems), Catania, Italy, pp. 342-356.

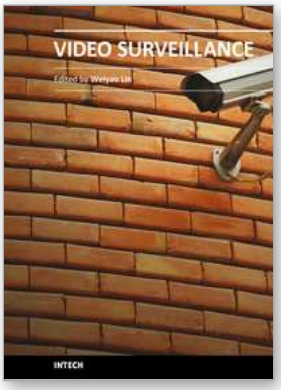
Smeaton, A., Over, P. & Kraaij, W. (2009). "High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements," in *Multimedia Content Analysis, Theory and Applicaitons*, Editor, Divakaran, A., pp. 151-174, Springer Verlag, ISBN: 978-0-387-76567-9, Berlin.

Stauffer, C. & Grimson, W. (1999). "Adaptive Background Mixture Models for Real-Time Tracking," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 246-252.

Szarvas, M., Sakai, U. & Ogata, J. (2006). "Real-time Pedestrian Detection Using LIDAR and Convolutional Neural Networks," *IEEE Intelligent Vehicles Symposium*, pp. 213-218.

Wang, L, Hu, W. & Tan, T. (2003). "Recent developments in human motion analysis," *Pattern Recognition* Vol. 36, No. 3, pp. 585-601.

Watson, A. & Ahumada, A. Jr. (1985). "Model of Human Visual-Motion sensing," *J. Opt. Soc. Am.*, A 2, pp. 322-342.



Video Surveillance

Edited by Prof. Weiyao Lin

ISBN 978-953-307-436-8
Hard cover, 486 pages
Publisher InTech
Published online 03, February, 2011
Published in print edition February, 2011

This book presents the latest achievements and developments in the field of video surveillance. The chapters selected for this book comprise a cross-section of topics that reflect a variety of perspectives and disciplinary backgrounds. Besides the introduction of new achievements in video surveillance, this book also presents some good overviews of the state-of-the-art technologies as well as some interesting advanced topics related to video surveillance. Summing up the wide range of issues presented in the book, it can be addressed to a quite broad audience, including both academic researchers and practitioners in halls of industries interested in scheduling theory and its applications. I believe this book can provide a clear picture of the current research status in the area of video surveillance and can also encourage the development of new achievements in this field.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821