A Survey on Visual Surveillance of Object Motion and Behaviors

Weiming Hu, Tieniu Tan, *Fellow, IEEE*, Liang Wang, and Steve Maybank

***Abstract*—Visual surveillance in dynamic scenes, especially for humans and vehicles, is currently one of the most active research topics in computer vision. It has a wide spectrum of promising applications, including access control in special areas, human identification at a distance, crowd flux statistics and congestion analysis, detection of anomalous behaviors, and interactive surveillance using multiple cameras, etc. In general, the processing framework of visual surveillance in dynamic scenes includes the following stages: modeling of environments, detection of motion, classification of moving objects, tracking, understanding and description of behaviors, human identification, and fusion of data from multiple cameras. We review recent developments and general strategies of all these stages. Finally, we analyze possible research directions, e.g., occlusion handling, a combination of two- and three-dimensional tracking, a combination of motion analysis and biometrics, anomaly detection and behavior prediction, content-based retrieval of surveillance videos, behavior understanding and natural language description, fusion of information from multiple sensors, and remote surveillance.**

***Index Terms*—Behavior understanding and description, fusion of data from multiple cameras, motion detection, personal identification, tracking, visual surveillance.**

I. Introduction

AS AN ACTIVE research topic in computer vision, visual surveillance in dynamic scenes attempts to detect, recognize and track certain objects from image sequences, and more generally to understand and describe object behaviors. The aim is to develop intelligent visual surveillance to replace the traditional passive video surveillance that is proving ineffective as the number of cameras exceeds the capability of human operators to monitor them. In short, 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.

Visual surveillance in dynamic scenes has a wide range of potential applications, such as a security guard for communities and important buildings, traffic surveillance in cities and expressways, detection of military targets, etc. We focus in this paper on applications involving the surveillance of people or vehicles, as they are typical of surveillance applications in general, and include the full range of surveillance methods. Surveillance applications involving people or vehicles include the following.

1) **Access control in special areas**. In some security-sensitive locations such as military bases and important governmental units, only people with a special identity are allowed to enter. A biometric feature database including legal visitors is built beforehand using biometric techniques. When somebody is about to enter, the system could automatically obtain the visitor's features, such as height, facial appearance and walking gait from images taken in real time, and then decide whether the visitor can be cleared for entry.
2) **Person-specific identification in certain scenes**. Personal identification at a distance by a smart surveillance system can help the police to catch suspects. The police may build a biometric feature database of suspects, and place visual surveillance systems at locations where the suspects usually appear, e.g., subway stations, casinos, etc. The systems automatically recognize and judge whether or not the people in view are suspects. If yes, alarms are given immediately. Such systems with face recognition have already been used at public sites, but the reliability is too low for police requirements.
3) **Crowd flux statistics and congestion analysis**. Using techniques for human detection, visual surveillance systems can automatically compute the flux of people at important public areas such as stores and travel sites, and then provide congestion analysis to assist in the management of the people. In the same way, visual surveillance systems can monitor expressways and junctions of the road network, and further analyze the traffic flow and the status of road congestion, which are of great importance for traffic management.
4) **Anomaly detection and alarming**. In some circumstances, it is necessary to analyze the behaviors of people and vehicles and determine whether these behaviors are normal or abnormal. For example, visual surveillance systems set in parking lots and supermarkets could analyze abnormal behaviors indicative of theft. Normally, there are two ways of giving an alarm. One way is to automatically make a recorded public announcement whenever any abnormal behavior is detected. The other is to contact the police automatically.
5) **Interactive surveillance using multiple cameras**. For social security, cooperative surveillance using multiple

Manuscript received April 14, 2003; revised September 26, 2003 and January 8, 2004. This work was supported in part by the National Science Foundation of China (NSFC) under Grants 60105002, 60335010, 60121302, and 60373046, by the Natural Science Foundation of Beijing under Grant 4031004 and Grant 4041004, by the National 863 High-Tech R&D Program of China under Grant 2002AA117010-11 and Grant 2002AA142100, and by the International Cooperation Project of Beijing, the LIAMA Project. This paper was recommended by Associate Editor D. Zhang.

W. Hu, T. Tan, and L. Wang are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China (e-mail: wmhu@nlpr.ia.ac.cn; tnt@nlpr.ia.ac.cn; lwang@nlpr.ia.ac.cn).

S. Maybank is with the School of Computer Science and Information Systems, Birkbeck College, London WC1E 7HX, U.K. (e-mail: sjmaybank@dcs.bbk.ac.uk).

Digital Object Identifier 10.1109/TSMCC.2004.829274

cameras could be used to ensure the security of an entire community, for example by tracking suspects over a wide area by using the cooperation of multiple cameras. For traffic management, interactive surveillance using multiple cameras can help the traffic police discover, track, and catch vehicles involved in traffic offences.

It is the broad range of applications that motivates the interests of researchers worldwide. For example, the IEEE has sponsored the IEEE International Workshop on Visual Surveillance on three occasions, in India (1998), the U.S. (1999), and Ireland (2000). In [68] and [1], a special section on visual surveillance was published in June and August of 2000, respectively. In [78], a special issue on visual analysis of human motion was published in March 2001. In [69], a special issue on third-generation surveillance systems was published in October 2001. In [130], a special issue on understanding visual behavior was published in October 2002. Recent developments in human motion analysis are briefly introduced in our previous paper [75]. It is noticeable that, after the 9/11 event, visual surveillance has received more attention not only from the academic community, but also from industry and governments.

Visual surveillance has been investigated worldwide under several large research projects. For example, the Defense Advanced Research Projection Agency (DARPA) supported the Visual Surveillance and Monitoring (VSAM) project [3] 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. Furthermore, to enhance protection from terrorist attacks, the Human Identification at a Distance (HID) program sponsored by DARPA in 2000 aims to develop a full range of multimodal surveillance technologies for successfully detecting, classifying, and identifying humans at great distances. The European Union's Framework V Programme sponsored Advisor, a core project on visual surveillance in metrostations.

There have been a number of famous visual surveillance systems. The real-time visual surveillance system W4 [4] 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. This system uses the single camera and grayscale sensor. The VIEWS system [87] at the University of Reading is a three-dimensional (3-D) model based vehicle tracking system. The Pfinder system developed by Wren *et al.* [8] is used to recover a 3-D description of a person in a large room. It tracks a single nonoccluded person in complex scenes, and has been used in many applications. As a single-person tracking system, TI, developed by Olsen *et al.* [9], detects moving objects in indoor scenes using motion detection, tracks them using first-order prediction, and recognizes behaviors by applying predicates to a graph formed by linking corresponding objects in successive frames. This system cannot handle small motions of background objects. The system at CMU [10] can monitor activities over a large area using multiple cameras that are connected into a network. It can detect and track multiple persons and vehicles within cluttered scenes and monitor their activities over long periods of time. The above comments on [8]–[10] are derived from [4]. Please see [4] for more details.

As far as hardware is concerned, companies like Sony and Intel have designed equipment suitable for visual surveillance, e.g., active cameras, smart cameras [76], omni-directional cameras [23], [77], etc.

All of the above activities are evidence of a great and growing interest in visual surveillance in dynamic scenes. The primary purpose of this paper is to give a general review on the overall process of a visual surveillance system. Fig. 1 shows the general framework of visual surveillance in dynamic scenes. The prerequisites for effective automatic surveillance using a single camera include the following stages: modeling of environments, detection of motion, classification of moving objects, tracking, understanding and description of behaviors, and human identification. In order to extend the surveillance area and overcome occlusion, fusion of data from multiple cameras is needed. This fusion can involve all the above stages. In this paper we review recent developments and analyze future open directions in visual surveillance in dynamic scenes. The main contributions of this paper are as follows.

- Low-level vision, intermediate-level vision, and high-level vision are discussed in a clearly organized hierarchical manner according to the general framework of visual surveillance. This, we believe, can help readers, especially newcomers to this area, not only to obtain an understanding of the state-of-the-art in visual surveillance, but also to appreciate the major components of a visual surveillance system and their inter-component links.
- Instead of detailed summaries of individual publications, our emphasis is on discussing various methods for different tasks involved in a general visual surveillance system. Each issue is accordingly divided into subprocesses or categories of various methods to examine the state of the art. Only the principles of each group of methods are described. The merits and demerits of a variety of different algorithms, especially for motion detection and tracking, are summarized.
- We give a detailed review of the state of the art in personal identification at a distance and fusion of data from multiple cameras.
- We provide detailed discussions on future research directions in visual surveillance, e.g., occlusion handling, combination of two-dimensional (2-D) tracking and 3-D tracking, combination of motion analysis and biometrics, anomaly detection and behavior prediction, behavior understanding and nature language description, content-based retrieval of surveillance videos, fusion of information from multiple sensors, and remote surveillance.

The remainder of this paper is organized as follows. Section II reviews the work related to motion detection including modeling of environments, segmentation of motion, classification of moving objects. Section III discusses tracking of objects, and Section IV details understanding and description of behaviors. Sections V and VI cover, respectively, personal identification at

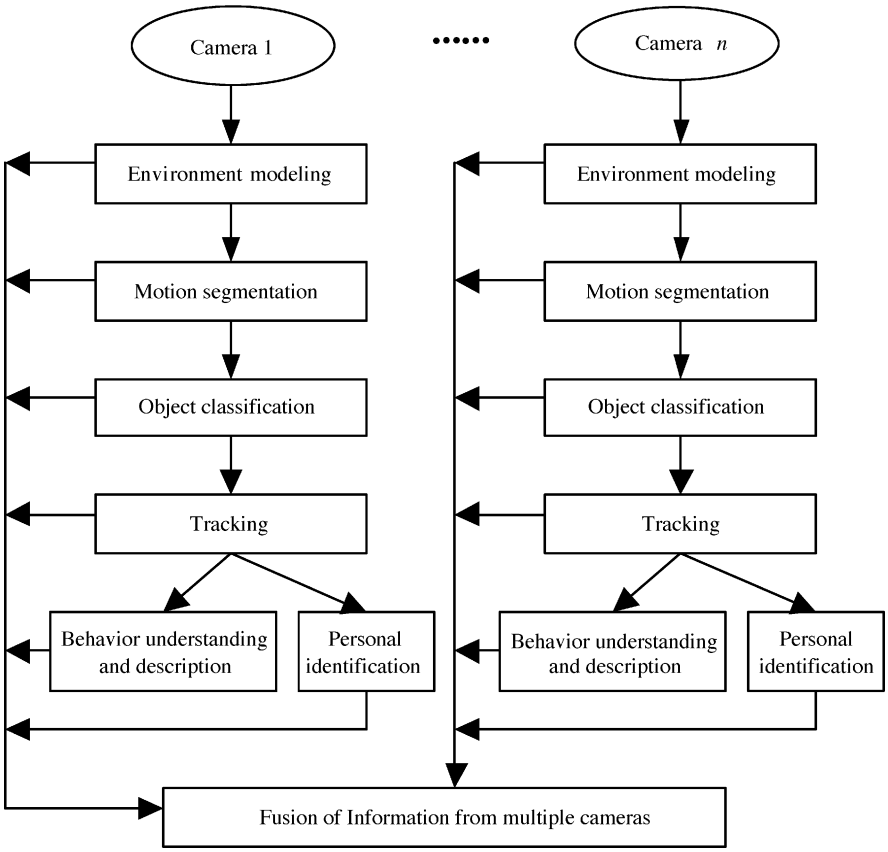


Fig. 1. General framework of visual surveillance.

a distance and fusion of data from multiple cameras. Section VII analyzes some possible directions for future research. The last section summarizes the paper.

II. Motion Detection

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

A. Environment Modeling

The active construction and updating of environmental models are indispensable to visual surveillance. Environmental models can be classified into 2-D models in the image plane and 3-D models in real world coordinates. Due to their simplicity, 2-D models have more applications.

- For fixed cameras, the key problem is to automatically recover and update background images from a dynamic sequence. Unfavorable factors, such as illumination variance, shadows and shaking branches, bring many difficulties to the acquirement and updating of background images. There are many algorithms for resolving these problems including temporal average of an image sequence [15], [82], adaptive Gaussian estimation [70], and parameter estimation based on pixel processes [79], [80], etc. Ridder *et al.* [81] model each pixel value with a Kalman Filter to compensate for illumination variance. Stauffer *et al.* [12], [80] present a theoretic framework for recovering and updating background images based on a process in which a mixed Gaussian model is used for each pixel value and online estimation is used to update background images in order to adapt to illumination variance and disturbance in backgrounds. Toyama *et al.* [83] propose the Wallflower algorithm in which background maintenance and background subtraction are carried out at three levels: the pixel level, the region level, and the frame level. Haritaoglu *et al.* [4] build a statistical model by representing each pixel with three values: its minimum and maximum intensity values, and the maximum intensity difference between consecutive frames observed during the training period. These three values are updated periodically. McKenna *et al.* [11] use an adaptive background model with color and gradient information to reduce the influences of shadows and unreliable color cues.
- For pure translation (PT) cameras, an environment model can be made by patching up a panorama graph to acquire a holistic background image [84]. Homography matrices can be used to describe the transformation relationship between different images.
- For mobile cameras, motion compensation is needed to construct temporary background images [85].

Regarding 3-D environmental models [86], current work is still limited to indoor scenes because of the difficulty of 3-D reconstructions of outdoor scenes.

B. Motion Segmentation

Motion segmentation in image sequences aims at detecting regions corresponding to moving objects such as vehicles and humans. 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. At present, most segmentation methods use either temporal or spatial information in the image sequence. Several conventional approaches for motion segmentation are outlined in the following.

1) **Background subtraction**. Background subtraction is a popular method for motion segmentation, especially under those situations with a relatively static background. It detects moving regions in an image by taking the difference between the current image and the reference background image in a pixel-by-pixel fashion. It is simple, but extremely sensitive to changes in dynamic scenes derived from lighting and extraneous events etc. Therefore, it is highly dependent on a good background model to reduce the influence of these changes [4], [11], [12], as part of environment modeling.
2) **Temporal differencing**. Temporal differencing makes use of the pixel-wise differences between two or three consecutive frames in an image sequence to extract moving regions. Temporal differencing is very 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. As an example of this method, Lipton *et al.* [10] detect moving targets in real video streams using temporal differencing. After the absolute difference between the current and the previous frame is obtained, a threshold function is used to determine changes. By using a connected component analysis, the extracted moving sections are clustered into motion regions. An improved version uses three-frame instead of two-frame differencing.
3) **Optical flow**. Optical-flow-based motion segmentation uses characteristics of flow vectors of moving objects over time to detect moving regions in an image sequence. For example, Meyer *et al.* [13], [21] compute the displacement vector field to initialize a contour based tracking algorithm, called active rays, for the extraction of articulated objects. The results are used for gait analysis. 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 [14].

Of course, besides the basic methods described above, there are some other approaches for motion segmentation. Using the extended expectation maximization (EM) algorithm, Friedman *et al.* [15] implement a mixed Gaussian classification model for each pixel. This model classifies the pixel values into three separate predetermined distributions corresponding to background, foreground and shadow. It also updates the mixed component automatically for each class according to the likelihood of membership. Hence, slowly moving objects are handled perfectly, while shadows are eliminated much more effectively. VSAM [3] has successfully developed a hybrid algorithm for motion segmentation by combining an adaptive background subtraction algorithm with a three-frame differencing technique. This hybrid algorithm is very fast and surprisingly effective for detecting moving objects in image sequences. In addition, Stringa [16] proposes a novel morphological algorithm for detecting motion in scenes. This algorithm obtains stable segmentation results even under varying environmental conditions.

C. Object Classification

Different moving regions may correspond to different moving targets in natural scenes. For instance, the image sequences captured by surveillance cameras mounted in road traffic scenes probably include humans, vehicles and other moving objects such as flying birds and moving clouds, etc. To further track objects and analyze their behaviors, it is essential to correctly classify moving objects. Object classification can be considered as a standard pattern recognition issue. At present, there are two main categories of approaches for classifying moving objects.

1) **Shape-based classification**. Different descriptions of shape information of motion regions such as points, boxes, silhouettes and blobs are available for classifying moving objects. VASM [3] takes image blob dispersedness, image blob area, apparent aspect ratio of the blob bounding box, etc, as key features, and classifies moving-object blobs into four classes: single human, vehicles, human groups, and clutter, using a viewpoint-specific three-layer neural network classifier. Lipton *et al.* [10] use the dispersedness and area of image blobs as classification metrics to classify all moving-object blobs into humans, vehicles and clutter. Temporal consistency constraints are considered so as to make classification results more precise. Kuno *et al.* [17] use simple shape parameters of human silhouette patterns to separate humans from other moving objects.
2) **Motion-based classification**. In general, nonrigid articulated human motion shows a periodic property, so this has been used as a strong cue for classification of moving objects. Cutler *et al.* [18] describe a similarity-based technique to detect and analyze periodic motion. By tracking an interesting moving object, its self-similarity is computed as it evolves over time. As we know, for periodic motion, its self-similarity measure is also periodic. Therefore time-frequency analysis is applied to detect and characterize the periodic motion, and tracking and classification of moving objects are implemented using periodicity. In Lipton's work [19], residual flow is used to analyze rigidity and periodicity of moving objects. It is expected that rigid objects present little residual flow, whereas a nonrigid moving object such as a human being has a higher average residual flow and even display a periodic component. Based on this useful cue, human motion is distinguished from motion of other objects, such as vehicles.

The two common approaches mentioned above, namely shape-based and motion-based classification, can also be effectively combined for classification of moving objects. Furthermore, Stauffer [20] proposes a novel method based on a time co-occurrence matrix to hierarchically classify both objects and behaviors. It is expected that more precise classification results can be obtained by using extra features such as color and velocity.

III. Object Tracking

After motion 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. Useful mathematical tools for tracking include the Kalman filter, the Condensation algorithm, the dynamic Bayesian network, the geodesic method, etc. Tracking methods are divided into four major categories: region-based tracking, active-contour-based tracking, feature-based tracking, and model-based tracking. It should be pointed out that this classification is not absolute in that algorithms from different categories can be integrated together [169].

A. Region-Based Tracking

Region-based tracking algorithms track objects according to variations of the image regions corresponding to the moving objects. For these algorithms, the background image is maintained dynamically [90], [91], and motion regions are usually detected by subtracting the background from the current image. Wren *et al.* [8] explore the use of small blob features to track a single human in an indoor environment. In their work, a human body is considered as a combination of some blobs respectively representing various body parts such as head, torso and the four limbs. Meanwhile, both human body and background scene are modeled with Gaussian distributions of pixel values. Finally, the pixels belonging to the human body are assigned to the different body part's blobs using the log-likelihood measure. Therefore, by tracking each small blob, the moving human is successfully tracked. Recently, McKenna *et al.* [11] propose an adaptive background subtraction method in which color and gradient information are combined to cope with shadows and unreliable color cues in motion segmentation. Tracking is then performed at three levels of abstraction: regions, people, and groups. Each region has a bounding box and regions can merge and split. A human is composed of one or more regions grouped together under the condition of geometric structure constraints on the human body, and a human group consists of one or more people grouped together. Therefore, using the region tracker and the individual color appearance model, perfect tracking of multiple people is achieved, even during occlusion. As far as region-based vehicle tracking is concerned, there are some typical systems such as the CMS mobilizer system supported by the Federal Highway Administration (FHWA), at the Jet Propulsion Laboratory (JPL) [92], and the PATH system developed by the Berkeley group [93].

Although they work well in scenes containing only a few objects (such as highways), region-based tracking algorithms cannot reliably handle occlusion between objects. Furthermore, as these algorithms only obtain the tracking results at the region level and are essentially procedures for motion detection, the outline or 3-D pose of objects cannot be acquired. (The 3-D pose of an object consists of the position and orientation of the object). Accordingly, these algorithms cannot satisfy the requirements for surveillance against a cluttered background or with multiple moving objects.

B. Active Contour-Based Tracking

Active contour-based tracking algorithms track objects by representing their outlines as bounding contours and updating these contours dynamically in successive frames [6], [71], [72], [74]. These algorithms aim at directly extracting shapes of subjects and provide more effective descriptions of objects than region-based algorithms. Paragios *et al.* [30] detect and track multiple moving objects in image sequences using a geodesic active contour objective function and a level set formulation scheme. Peterfreund [31] explores a new active contour model based on a Kalman filter for tracking nonrigid moving targets such as people in spatio-velocity space. Isard *et al.* [32] adopt stochastic differential equations to describe complex motion models, and combine this approach with deformable templates to cope with people tracking. Malik *et al.* [82], [94] have successfully applied active contour-based methods to vehicle tracking.

In contrast to region-based tracking algorithms, active contour-based algorithms describe objects more simply and more effectively and reduce computational complexity. Even under disturbance or partial occlusion, these algorithms may track objects continuously. However, the tracking precision is limited at the contour level. The recovery of the 3-D pose of an object from its contour on the image plane is a demanding problem. A further difficulty is that the active contour-based algorithms are highly sensitive to the initialization of tracking, making it difficult to start tracking automatically.

C. Feature-Based Tracking

Feature-based tracking algorithms perform recognition and tracking of objects by extracting elements, clustering them into higher level features and then matching the features between images. Feature-based tracking algorithms can further be classified into three subcategories according to the nature of selected features: global feature-based algorithms, local feature-based algorithms, and dependence-graph-based algorithms.

- The features used in global feature-based algorithms include centroids, perimeters, areas, some orders of quadratures and colors [100], [101], etc. Polana *et al.* [33] provide a good example of global feature-based tracking. A person is bounded with a rectangular box whose centroid is selected as the feature for tracking. Even when occlusion happens between two persons during tracking, as long as the velocity of the centroids can be distinguished effectively, tracking is still successful.
- The features used in local feature-based algorithms include line segments, curve segments, and corner vertices [98], [99], etc.

- The features used in dependence-graph-based algorithms include a variety of distances and geometric relations between features [97].

The above three methods can be combined. In the recent work of Jang *et al.* [34], an active template that characterizes regional and structural features of an object is built dynamically based on the information of shape, texture, color, and edge features of the region. Using motion estimation based on a Kalman filter, the tracking of a nonrigid moving object is successfully performed by minimizing a feature energy function during the matching process.

In general, as they operate on 2-D image planes, feature-based tracking algorithms can adapt successfully and rapidly to allow real-time processing and tracking of multiple objects which are required in heavy thruway scenes, etc. However, dependence-graph-based algorithms cannot be used in real-time tracking because they need time-consuming searching and matching of graphs. Feature-based tracking algorithms can handle partial occlusion by using information on object motion, local features and dependence graphs. However, there are several serious deficiencies in feature-based tracking algorithms.

- The recognition rate of objects based on 2-D image features is low, because of the nonlinear distortion during perspective projection and the image variations with the viewpoint's movement.
- These algorithms are generally unable to recover 3-D pose of objects.
- The stability of dealing effectively with occlusion, overlapping and interference of unrelated structures is generally poor.

D. Model-Based Tracking

Model-based tracking algorithms track objects by matching projected object models, produced with prior knowledge, to image data. The models are usually constructed off-line with manual measurement, CAD tools or computer vision techniques. As model-based rigid object tracking and model-based nonrigid object tracking are quite different, we review separately model-based human body tracking (nonrigid object tracking) and model-based vehicle tracking (rigid object tracking).

1) Model-Based Human Body Tracking: The general approach for model-based human body tracing is known as analysis-by-synthesis, and it is used in a predict-match-update style. Firstly, the pose of the model for the next frame is predicted according to prior knowledge and tracking history. Then, the predicted model is synthesized and projected into the image plane for comparison with the image data. A specific pose evaluation function is needed to measure the similarity between the projected model and the image data. According to different search strategies, this is done either recursively or using sampling techniques until the correct pose is finally found and is used to update the model. Pose estimation in the first frame needs to be handled specially. Generally, model-based human body tracking involves three main issues:

- construction of human body models;
- representation of prior knowledge of motion models and motion constraints;
- prediction and search strategies.

Previous work on these three issues is briefly and respectively reviewed as follows.

a) Human body models: Construction of human body models is the base of model-based human body tracking [24]. Generally, the more complex a human body model, the more accurate the tracking results, but the more expensive the computation. Traditionally, the geometric structure of human body can be represented in the following four styles.

- **Stick figure**. The essence of human motion is typically contained in the movements of the torso, the head and the four limbs, so the stick-figure method is to represent the parts of a human body as sticks and link the sticks with joints. Karaulova *et al.* [25] use a stick figure representation to build a novel hierarchical model of human dynamics encoded using hidden Markov models (HMMs), and realize view-independent tracking of a human body in monocular image sequences.
- **2-D contour**. This kind of human body model is directly relevant to human body projections in an image plane. The human body segments are modeled by 2-D ribbons or blobs. For instance, Ju *et al.* [26] propose a cardboard human body model, in which the human limbs are represented by a set of jointed planar ribbons. The parameterized image motion of these patches is constrained to enforce the articulated movement of human limbs. Niyogi *et al.* [27] use the spatial-temporal pattern in XYT space to track, analyze and recognize walking figures. They examine the characteristic braided pattern produced by the lower limbs of a walking human, the projections of head movements are then located in the spatio-temporal domain, followed by the identification of the joint trajectories; The contour of a walking figure is outlined by utilizing these joint trajectories, and a more accurate gait analysis is carried out using the outlined 2-D contour for the recognition of the specific human.
- **Volumetric models**. The main disadvantage of 2-D models is that they require restrictions on the viewing angle. To overcome this disadvantage, many researchers use 3-D volumetric models such as elliptical cylinders, cones [102], [103], spheres, super-quadrics [104], etc. Volumetric models require more parameters than image-based models and lead to more expensive computation during the matching process. Rohr [28] makes use of fourteen elliptical cylinders to model a human body in 3-D volumes. Wachter *et al.* [29] establish a 3-D body model using connected elliptical cones.
- **Hierarchical model**. Plankers *et al.* [105] present a hierarchical human model for achieving more accurate results. It includes four levels: skeleton, ellipsoid meatballs simulating tissues and fats, polygonal surface representing skin, and shaded rendering.

b) Motion models: Motion models of human limbs and joints are widely used in tracking. They are effective because the movements of the limbs are strongly constrained. These motion models serve as prior knowledge to predict motion parameters [106], [107], to interpret and recognize human behaviors [108], or to constrain the estimation of low-level image measurements

[109]. For instance, Bregler [108] decomposes a human behavior into multiple abstractions, and represents the high-level abstraction by HMMs built from phases of simple movements. This representation is used for both tracking and recognition. Zhao *et al.* [106] learn a highly structured motion model for ballet dancing under the minimum description length (MDL) paradigm. This motion model is similar to a finite-state machine (FSM). The multivariate principal component analysis (MPCA) is used to train a walking model in Sidenbladh *et al.*'s work [109]. Similarly, Ong *et al.* [110] employ the hierarchical PCA to learn their motion model which is based on the matrices of transition probabilities between different subspaces in a global eigensapce and the matrix of transition probabilities between global eigenspaces. Ning *et al.* [7] learn a motion model from semi-automatically acquired training examples and represent it using Gaussian distributions.

c) Search strategies: Pose estimation in a high-dimensional body configuration space is intrinsically difficult, so, search strategies are often carefully designed to reduce the solution space. Generally, there are four main classes of search strategies: dynamics, Taylor models, Kalman filtering, and stochastic sampling. Dynamical strategies use physical forces applied to each rigid part of the 3-D model of the tracked object. These forces, as heuristic information, guide the minimization of the difference between the pose of the 3-D model and the pose of the real object [102]. The strategy based on the Taylor models incrementally improves an existing estimation, using differentials of motion parameters with respect to the observation to predict better search directions [112]. It at least finds local minima, but cannot guarantee finding the global minimum. As a recursive linear estimator, Kalman filtering can thoroughly deal with the tracking of shape and position over time in the relatively clutter-free case in which the density of the motion parameters can be modeled satisfactorily as Gaussian [29], [114]. To handle clutter that causes the probability density function for motion parameters to be multimodal and non-Gaussian, stochastic sampling strategies, such as Markov Chain Monte Carlo [115], Genetic Algorithms, and CONDENSATION [116], [117], are designed to represent simultaneous alternative hypotheses. Among the stochastic sampling strategies in visual tracking, CONDENSATION is perhaps the most popular.

2) Model-Based Vehicle Tracking: As to model-based vehicle tracking, 3-D wire-frame vehicle models are mainly used [95]. The research groups at the University of Reading [87], [88], the National Laboratory of Pattern Recognition (NLPR) [111], [174] and the University of Karlsruhe [124]–[126] have made important contributes to 3-D model-based vehicle localization and tracking.

The research group at the University of Reading adopts 3-D wire-frame vehicle models. Tan *et al.* [87], [119] propose the ground-plane constraint (GPC), under which vehicles are restricted to move on the ground plane. Thus the degrees of freedom of vehicle pose are reduced to three from six. This greatly decreases the computational cost of searching for the optimal pose. Moreover, under the weak perspective assumption, the pose parameters are decomposed into two independent sets: translation parameters and rotation parameters. Tan *et al.* [120] propose a generalized Hough transformation algorithm based on a single characteristic line segment matching to estimate vehicle pose. Further, Tan *et al.* [121] analyze the one-dimensional (1-D) correlation of image gradients and determine the vehicle pose by voting. As to the refinement of the vehicle pose, the research group in the University of Reading have utilized an independent 1-D searching method [121] in their past work. Recently, Pece *et al.* [122], [123] introduce a statistical Newton method for estimating the vehicle pose.

The NLPR group has extended the work of the research group at the University of Reading. Yang *et al.* [111] propose a new 3-D model-based vehicle localization algorithm, in which the edge points in the image are directly used as features, and the degree of matching between the edge points and the projected model is measured by a pose evaluation function. Lou *et al.* [174] present an algorithm for vehicle tracking based on an improved extended Kalman filter. In the algorithm, the turn of the steering wheel and the distance between the front and rear wheels are taken into account. As there is a direct link between the behavior of the driver who controls the motion of the vehicle and the assumed dynamic model, the improved extended Kalman filter outperforms the traditional extended Kalman filter when the vehicle carries out a complicated maneuver.

The Karlsruhe group [124] uses the 3-D wire-frame vehicle model. The image features used in the algorithm are edges. The initial values for the vehicle pose parameters are obtained from the correspondence between the segments in an image and those in the projection model. The correspondence is calculated using viewpoint consistent constraints and some clustering rules. The *maximum a posterior* (MAP) estimate of the vehicle position is obtained using the Levenberg–Marquardt optimization technique. The algorithm is data-driven and dependent on the accuracy of edge detection. Kollnig *et al.* [125] also propose an algorithm based on image gradients, in which virtual gradients in an image are produced by spreading the Gaussian distribution around line segments. Under the assumption that the real gradient at each point in an image is the sum of a virtual gradient and a Gaussian white noise, the pose parameters can be estimated using the extended Kalman filter (EKF). Furthermore, Haag *et al.* [126] integrate Kollnig *et al.*'s algorithm based on image gradients with that based on optic flow. The method uses image gradients evaluated in the neighborhoods of the image features. However, the optic flow uses global information on image features, integrated across the whole region of interest (ROI). So the gradients and the optic flow are complementary sources of information.

The above reviews model-based human body tracking and model-based vehicle tracking. Compared with other tracking algorithms, model-based tracking algorithms have the following main advantages.

- By making use of the prior knowledge of the 3-D contours or surfaces of objects, the algorithms are intrinsically robust. The algorithms can obtain better results even under occlusion (including self-occlusion for humans) or interference between nearby image motions.
- As far as model-based human body tracking is concerned, the structure of human body, the constraint of human motion, and other prior knowledge can be fused.

- As far as 3-D model-based tracking is concerned, after setting up the geometric correspondence between 2-D image coordinates and 3-D world coordinates by camera calibration, the algorithms naturally acquire the 3-D pose of objects.
- The 3-D model-based tracking algorithms can be applied even when objects greatly change their orientations during the motion.

Ineluctably, model-based tracking algorithms have some disadvantages such as the necessity of constructing the models, high computational cost, etc.

IV. Understanding and Description of Behaviors

After successfully tracking the moving objects from one frame to another in an image sequence, the problem of understanding-object behaviors from image sequences follows naturally. Behavior understanding involves the analysis and recognition of motion patterns, and the production of high-level description of actions and interactions.

A. Behavior Understanding

Understanding of behaviors may simply be thought as the classification of time varying feature data, i.e., matching an unknown test sequence with a group of labeled reference sequences representing typical behaviors. It is then obvious that a 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 patterns. Some efforts have been made in this direction [176] and the major existing methods for behavior understanding are outlined in the following.

a) Dynamic time warping (DTW): 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 [127], [128]. For instance, Bobick *et al.* [128] 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) Finite-state machine (FSM): 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. Wilson *et al.* [129] analyze the explicit structure of natural gestures where the structure is implemented by an equivalent of a FSM but with no learning involved. State-machine representations of behaviors have also been employed in higher level description. For instance, Bremond *et al.* [131] use handcrafted deterministic automata to recognize airborne surveillance scenarios describing vehicle behaviors in aerial imagery.

c) HMMs: A HMM is a kind of stochastic state machines [35]. It allows a more sophisticated analysis of data with spatio-temporal variability. The use of HMMs consists of two stages: training and classification. In the training stage, 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. In the matching stage, the probability with which a particular HMM generates the test symbol sequence corresponding to the observed image features is computed. HMMs generally outperform DTW for undivided time series data, and are therefore extensively applied to behavior understanding. For instance, Starner *et al.* [132] propose HMMs for the recognition of sign language. Oliver *et al.* [133] propose and compare two different state-based learning architectures, namely, HMMs and coupled hidden Markov models (CHMMs) for modeling people behaviors and interactions such as following and meeting. The CHMMs are shown to work much more efficiently and accurately than HMMs. Brand *et al.* [134] show that, by the use of the entropy of the joint distribution to learn the HMM, a HMM's internal state machine can be made to organize observed behaviors into meaningful states. This technique has found applications in video monitoring and annotation, in low bit-rate coding of scene behaviors, and in anomaly 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 [135] and lip-reading [136].

e) Syntactic techniques [137]: The syntactic approach in machine vision has been studied mostly in the context of pattern recognition in static images. Recently the grammatical approach has been used for visual behavior recognition. Brand [138] uses a simple nonprobabilistic grammar to recognize sequences of discrete behaviors. Ivanov *et al.* [137] describe a probabilistic syntactic approach to the detection and recognition of temporally extended behaviors and interactions between multiple agents. The fundamental idea 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 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 behaviors in a given domain.

f) Non-deterministic finite automaton (NFA): Wada *et al.* [139] employ NFA as a sequence analyzer, because it is a simple example satisfying the following properties: instantaneousness and pure-nondeterminism. They present an approach for multi-object behavior recognition based on behavior driven selective attention.

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. Johnson *et al.* [140] describe the movement of an object in terms of a sequence of flow vectors, each of which consists of 4 components representing the positions and velocities of the object in the image plane. A statistical model of object trajectories is formed with two competitive learning networks that are connected with leaky neurons. Sumpter *et al.* [141] introduce feedback to the second competitive network in [140] giving a more efficient prediction of object behaviors. Hu *et al.* [175] improve the work in [140] by introducing a new neural network structure that has smaller scale and faster learning speed, and is thus more effective. Owens *et al.* [142] apply the Kohonen self-organizing feature map to find the flow vector distribution patterns. These patterns are used to determine whether a point on a trajectory is normal or abnormal.

B. Natural Language Description of Behaviors

In many applications it is important to describe object behaviors in natural language suitable for nonspecialist operator of visual surveillance [22], [147]. For example, Herzog *et al.* [143] have developed the VITRA project that uses natural language to describe visual scenes. In 1995, MIT [147] convened a workshop to discuss how to link natural language and computer vision. Generally, there are two main categories of behavior description methods: statistical models and formalized reasoning.

a) Statistical models: A representative statistical model is the Bayesian network model [144], [145]. This model interprets certain events and behaviors by analysis of time sequences and statistical modeling. For example, Remagnino *et al.* [148] describe interactions between objects using a two-layer agent-based Bayesian network. These methods rest on lower-level recognition based on motion concepts, and do not yet involve high-level concepts, such as events and scenarios, and the relationships between these concepts. These concepts need high-level reasoning based on a large amount of prior knowledge.

b) Formalized reasoning: Formalized reasoning [146] uses symbol systems to represent behavior patterns, and reasoning methods such as predication logic to recognize and classify events. Recently, Kojima *et al.* [36], [37] propose a new method for generating natural language descriptions of human behaviors appearing in real image sequences. First, a head region of a human is extracted from each image frame, and the 3-D pose and position of the head are estimated using a model-based approach. Next, the head motion trajectory is divided into the segments of monotonous movement. The conceptual features for each segment, such as degrees of changes of pose and position and the relative distances from other objects in the surroundings, are evaluated. Meanwhile, the most suitable verbs and other syntactic elements are selected. Finally, the natural language text for interpreting human behaviors is generated by machine translation technology. Kollnig *et al.* [118] use fuzzy membership functions to associate verbs with quantitative details obtained by automatic image sequence analysis for generating natural language descriptions of a traffic scene. In their scheme, each occurrence is defined by three predicates: a precondition, monotonicity condition and post-condition. The most significant disadvantage of the formalized reasoning methods is that they cannot handle uncertainty of events [96].

Although there is some progress in description of behaviors, some key problems remain open, for example how to properly represent semantic concepts, how to map motion characteristics to semantic concepts and how to choose efficient representations to interpret the scene meanings.

V. Personal Identification for Visual Surveillance

The problem of "who is now entering the area under surveillance" is of increasing importance for visual surveillance. Such personal identification can be treated as a special behavior-understanding problem. Human face and gait are now regarded as the main biometric features that can be used for personal identification in visual surveillance systems [2]. In recent years, great progress in face recognition [46]–[50] has been achieved. The main steps in the face recognition for visual surveillance are face detection, face tracking, face feature detection and face recognition [51]–[55]. Usually, these steps are studied separately. Therefore, developing an integrated face recognition system involving all of the above steps is critical for visual surveillance. As the length of this paper is restricted, we review here only recent researches on the major existing methods for gait recognition.

A. Model-Based Methods

In model-based methods, parameters, such as joint trajectories, limb lengths, and angular speeds, are measured [156]–[162], [180], [181].

Cunado *et al.* [156], [157] model gait as the movement of an articulated pendulum and use the dynamic Hough transform [158] to extract the lines representing the thigh in each frame. The least squares method is used to smooth the inclination data of the thigh and to fill the missing points caused by self-occlusion of the legs. Phase-weighted magnitude spectra are used as gait features for recognition.

Yam *et al.* [159], [160] propose a new model-based gait recognition algorithm. Biomechanical models of walking and running are used to form a type of new anatomical model called a dynamically coupled oscillator, for the hip motion, and the structure and motion of the thigh and the lower leg. Temporal template matching [161] is used to extract the rotation angles of the thigh and the lower legs. Then gait signatures are obtained from the lower-order phase-weighted magnitude spectra.

Another recent paper [162] uses dynamic features from trajectories of lower-body joint angles such as the hip and the knee to recognize individuals. This work first projects the 3-D positions of markers attached to the body into the walking plane. Then a simple method is applied to estimate the planar offsets between the markers and the underlying skeleton or joints. Finally, given these offsets, the trajectories of joint angles are computed.

Tracking and localizing the human body accurately in 3-D space is still difficult despite the recent work on structure-based methods. In theory, joint angles are sufficient for recognition of people by their gait. However, accurately recovering joint angles

from a walking video is still an unsolved or not well-solved problem. In addition, the computational cost of the model-based approaches is quite high.

B. Statistical Methods

Statistical recognition techniques usually characterize the statistical description of motion image sets, and have been well developed in automatic gait recognition [60], [163]–[168], [182], [184].

Murase *et al.* [163] use a parametric eigenspace representation to reduce computational cost and to improve the robustness of gait estimation. Huang *et al.* [164]–[166] have successfully extended Murase *et al.*'s work by adding canonical space analysis to obtain better discrimination. The eigenspace transformation (EST) has the advantage of reducing the dimensionality, but it cannot optimize class discrimination. Therefore, Huang *et al.* [165] describe an integrated gait recognition system using EST and canonical space analysis (CSA).

Shutler *et al.* [167] develop a velocity-moment-based method for describing the object motion in image sequences. Similarly, Lee *et al.* [60], [168] use the moment features of image regions to recognize individuals. Assuming that people walk frontal-parallel toward a fixed camera, the silhouette region is divided into seven subregions. A set of moment-based region features is used to recognize people and to predict the gender of an unknown person by his/her walking appearance.

Statistical methods are relatively robust to noise and change of time interval in input image sequences. Compared with model-based approaches, the computational cost of statistical methods is low.

C. Physical-Parameter-Based Methods

Physical-parameter-based methods make use of geometric structural properties of a human body to characterize a person's gait pattern. The parameters used include height, weight, stride cadence and length, etc. [56], [170]–[173], [183].

For example, a gait recognition technique using specific behavior parameters is recently proposed by Bobick *et al.* [170], [171]. This method does not directly analyze the dynamics of gait patterns, but uses walking activities to recover the static body parameters of walking such as the vertical distance between head and feet, the distance between head and pelvis, the distance between feet and pelvis, and the distance between the left and right feet. The method is assessed using an expected confusion metric [172] to predict how well a given feature vector can identify an individual in a large population.

Some recent work [56], [173] also uses human stature, stride length and cadence as the input features for parametric gait classification. Given the calibration parameters of the camera and the walking plane, the method uses the walking periodicity to accurately estimate cadence and stride [56].

Physical-parameter-based methods are intuitively understandable, and independent of viewing angles because these parameters usually are recovered in the 3-D space. However, they depend greatly on the vision techniques used to recover the required parameters, e.g., body-part labeling, depth compensation, camera calibration, shadow removal, etc. In addition, the parameters used for recognition may be not effective enough across a large population.

D. Spatio-Temporal Motion-Based Methods

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. These methods generally consider motion as a whole to characterize its spatio-temporal distributions [27], [58], [59], [61], [62], [177], [178], [185].

Perhaps the earliest approach to recognizing people is to obtain gait features from the spatio-temporal pattern of a walking figure [27]. In translation and time (XT) space, the motions of the head and legs have different patterns. These patterns are first processed to determine the bounding box of a moving body, and then fitted to a five-stick model. Gait signatures could be acquired from the velocity-normalized fitted model. Later, Niyogi *et al.* [58] extend their own work by using the spatio-temporal surface to analyze gait. After motion detection, the XYT pattern (2-D space and 1-D time) is fitted with a smooth spatio-temporal surface. This surface is represented as a combination of a standard parametric surface and a difference surface that can be used to recognize some simple actions.

Using the image self-similarity in XYT, BenAbdelkader *et al.* [59], [177] propose a motion-based gait-recognition technique. The similarity plots (SPs) of the image sequence of a moving object are projections of its planar dynamics [61]. Hence, these SPs include much information of gait motion.

Kale *et al.* [62] propose a HMM-based method for representing and recognizing gait. First, a set of key frames that occur during a walk cycle is chosen. The widths of the walking figure's binary silhouettes, in such a set of key frames, are chosen as the input features. Then, a low-dimensional measurement vector is produced using the Euclidean distance between a given image and the set of key frames. These measurement vector sequences are used to train the HMMs.

Spatio-temporal motion-based methods are able to better capture both spatial and temporal information of gait motion. Their advantage is low computational complexity and a simple implementation. However, they are susceptible to noise and to variations of the timings of movements.

E. Fusion of Gait With Other Biometrics

The fusion of gait information with other biometrics can further increase recognition robustness and reliability. Shakhnarovich *et al.* [64] develop a view-normalized method for solving the problem of integrated face and gait recognition from multiple views. For optimal face recognition, they set a virtual camera to capture the frontal face. For gait recognition, they set a virtual camera to capture the side-view walking sequence. Results show that the integrated face and gait recognition outperforms recognition which only uses a single mode. In extended work, Shakhnarovich *et al.* [65] evaluate the recognition performances of several different probabilistic combinations for fusing view-normalized face and gait.

Although many researchers have been working on gait recognition, current research of gait recognition is still in its infancy.

First, most experiments are carried out under constrained circumstances, e.g., no occlusion happens while objects are usually moving, the background is simple, etc. Second, existing algorithms are evaluated on small databases. Future work on gait recognition will focus on handling these two problems.

VI. Fusion of Data From Multiple Cameras

Motion detection, tracking, behavior understanding, and personal identification at a distance discussed above 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 multicameras 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.

A. Installation

The deployment of the cameras has a great influence on the real-time performance and the cost of the system. Cameras cannot be employed arbitrarily due to factors such as the topography of the area. Redundant cameras increase not only processing time and algorithmic complexity, but also the installation cost. In contrast, a lack of cameras may cause some blind angles, which reduce the reliability of a surveillance system. So the question of how to cover the entire scene with the minimum number of cameras is important. Pavlidis *et al.* [149] provide an optimum algorithm for solving the problem of multiple-camera installation in parking lots. The basic idea is to place camera 1 on a certain position at first, then to search the rest of the space to place the second camera at a point where there is a 25%–50% overlap region between the fields of view of camera 1 and camera 2. The other cameras are added one by one, subject to the constraint that the field of view of each new camera should have a 25%–50% overlap with the combined fields of view of all the previous cameras.

B. Calibration

Traditional calibration methods use the 3-D coordinates and the image coordinates of some known points to compute the parameters of a camera. Calibration is more complex when multiple cameras are concerned. Current multiple camera self-calibration methods use temporal information. Stein and Lee *et al.* [38], [39] use the motion trajectory and the ground plane constraint to determine the projection transformation matrix, and then such matrix is decomposed to obtain the extrinsic parameters of the camera. However, this method is inaccurate, and cannot be used if there is no ground plane.

C. Object Matching

Object matching among multiple cameras involves finding the correspondences between the objects in different image sequences taken by different cameras. There are two popular methods: one is the geometry-based method that establishes correspondence according to geometric features transformed to the same space; and the other is the recognition-based method. As an example of the geometry-based method, Cai *et al.* [41], [42] use features of location, intensity and geometry to match between images taken by different cameras. As an example of recognition-based methods, Krumm *et al.* [57] use color histograms to match regions. In general, the methods for object matching need camera calibration. However, some researchers also develop methods without calibration. For example, Javed *et al.* [150] use the spatial relationships between view fields of cameras to establish the corresponding relationships of images.

D. Switching

When an object moves out of the view field of an active camera, or the camera cannot give a good view of the moving object, then the system should switch to another camera that may give a better view of the object. The key problems are how to find the better camera and how to minimize the number of switches during tracking. Cai *et al.* [41], [42] establish a tracking confidence for each object. When the tracking confidence is below a certain threshold, the system begins a global search and selects the camera with the highest tracking confidence as the active camera.

E. Data Fusion

Data fusion is important for occlusion handling and continuous tracking. Dockstader *et al.* [151] use a Bayesian network to fuse 2-D state vectors acquired from various image sequences to obtain a 3-D state vector. Collins *et al.* [152] design an algorithm that obtains an integrated representation of an entire scene by fusing information from every camera into a 3-D geometric coordinate system. Kettnaker *et al.* [43] synthesize the tracking results of different cameras to obtain an integrated trajectory.

F. Occlusion Handling

In practice, self-occlusion, and occlusions between different moving objects or between moving objects and the background are inevitable. Multiple camera systems offer efficient and promising methods for coping with occlusion. Utsumi *et al.* [40] utilize multiple cameras to track people, successfully resolving the mutual-occlusion and self-occlusion by choosing the "best" view. Dockstader *et al.* [151] describe a multiple camera surveillance system that is used to track partly occluded people. Tsutsui *et al.* [153] apply the multiple camera surveillance system to optical flow-based human tracking. When a static object in one camera occludes an object, the system predicts the 3-D coordinate position and moving speed of the occluded object according to information from other cameras. Mittal *et al.* [154] resolve human tracking in complex scenes using multiple cameras. First, using the Bayesian classification rule, images are segmented according to the human model and the estimated position of each person. Then, data from multiple cameras are fused to estimate the positions of the humans on the ground plane. Finally, a Kalman filter is used for tracking.

VII. Future Developments

In Sections II–VI, we have reviewed the state-of-the-art of visual surveillance for humans and vehicles sorted by a general framework of visual surveillance systems. Although a large amount of work has been done in visual surveillance for humans and vehicles, many issues are still open and deserve further research, especially in the following areas.

A. Occlusion Handling

Occlusion handing is a major problem in visual surveillance. Typically, during occlusion, only portions of each object are visible and often at very low resolution. This problem is generally intractable, and motion segmentation based on background subtraction may become unreliable. To reduce ambiguities due to occlusion, better models need be developed to cope with the correspondence between features and body parts, and thus eliminate correspondence errors that occur during tracking multiple objects. When objects are occluded by fixed objects such as buildings and street lamps, some resolution is possible through motion region analysis and partial matching. However, when multiple moving objects occlude each other, especially when their speeds, directions and shapes are very close, their motion regions coalesce, which makes the location and tracking of objects particularly difficult. The self-occlusion of a human body is also a significant and difficult problem. Interesting progress is being made using statistical methods to predict object pose, position, and so on, from available image information. Perhaps the most promising practical method for addressing occlusion is through the use of multiple cameras.

B. Fusion of 2-D and 3-D Tracking

Two-dimensional tracking is simple and rapid, and it has shown some early successes in visual surveillance, especially for low-resolution application areas where the precise posture reconstruction is not needed, e.g., pedestrian and vehicle tracking in a traffic surveillance setting. However, the major drawback of the 2-D approach is its restriction of the camera angle.

Compared with 2-D approaches, 3-D approaches are more effective for accurate estimation of position in space, more effective handling of occlusion, and high-level judgments about complex object movements such as wandering around, shaking hands, dancing, and vehicle overtaking. However, applying 3-D tracking requires more parameters and more computation during the matching process. Also, vision-based 3-D tracking brings a number of challenges such as the acquisition of object models, occlusion handling, parameterized object modeling, etc.

In fact, the combination of 2-D tracking and 3-D tracking is a significant research direction that few researches have attempted. This combination is expected to fuse the merits of the 2-D tracking algorithms and those of the 3-D tracking algorithms. The main difficulties of this combination are:

- deciding when 2-D tracking should be used and when 3-D tracking should be used;
- how to initialize pose parameters for 3-D tracking according to the results from 2-D tracking, when the tracking algorithm is switched from 2-D to 3-D.

C. Three-Dimensional Modeling of Humans and Vehicles

We think that it is feasible to build 3-D models for humans and vehicles. As far as vehicles are concerned, they can be treated as rigid objects, drawn from only a few classes and with invariable 3-D shapes during normal usage. It is possible to establish 3-D models of vehicles using CAD tools, etc. A generic and parametric model can be established for each class [125], [155]. As far as human beings are concerned, the shapes of human bodies are similar, so it is possible to build a uniform parametric model for human bodies. The parametric models and their applications in tracking and identification are important research directions in visual surveillance. 3-D modeling deserves more attention in future work.

D. Combination of Visual Surveillance and Personal Identification

As mentioned in Section V, vision-based human identification at a distance has become increasingly important. Gait is a most attractive modality used for this purpose. Generally, future work on gait recognition will focus on the following directions.

1) **Establishing a large common database and a standard test protocol**. The database with an independent subdatabase for the test just like the FERET protocol [113] is necessary for convincing test. Any realistic database should include the factors affecting gait perception, e.g., clothing, environments, distance, carried objects such as briefcases [179], and viewing angle. Such a database allows one to explore the limitations of the extracted gait signatures as well as the confidence estimation associated with the use of gait to buttress other biometric measures [66].
2) **Combining dynamic features and static features**. Gait includes both individual appearances and the dynamics of walking. Developing the underlying static parameters of a human body and the dynamic characteristics of joint angles is helpful to recognition.
3) **Developing multiple biometric feature-based systems in which gait is a basic biometric feature**. A multiple biometric system either fuses multiple biometric features or automatically switches among different biometric features according to operational conditions. For example, at a distance, gait can be used for recognition; when an individual is near to the camera, the face image provides a powerful cue; at intermediate distances, the information from both face and gait can be fused to improve the recognition accuracy.
4) **Obtaining the view-invariant gait signatures from the tracked image sequences** [67]. To extract and localize arbitrarily articulated shapes, view-invariant gait signatures from the tracked image sequences need to be obtained in future recognition systems.

E. Behavior Understanding

One of the objectives of visual surveillance is to analyze and interpret individual behaviors and interactions between objects to decide for example whether people are carrying, depositing or exchanging objects, whether people are getting on or getting off a vehicle, or whether a vehicle is overtaking another vehicle, etc. 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. 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. This ambiguity is exacerbated when several objects are present in a scene [130]. The following problems within behavior understanding are challenging: statistical learning for modeling behaviors, context-dependent learning from example images, real-time performance required by behavior interpretation, classification and labeling of motion trajectories of tracked objects, automated learning of the *priori* knowledge [63] implied in object behaviors, visually mediated interaction, and attention mechanisms.

F. Anomaly Detection and Behavior Prediction

Anomaly detection and behavior prediction are significant in practice. In applications of visual surveillance, not only should visual surveillance systems detect anomalies such as traffic accidents and car theft etc, according to requirements of functions, but also predict what will happen according to the current situation and raise an alarm for a predicted abnormal behavior. Implementations are usually based on one or other of the following two methods.

1) **Probability reasoning and prior rules combined methods**. A behavior with small probability, or against the prior rules would be regarded as an anomaly.
2) **Behavior-pattern-based methods**. Based on learned patterns of behaviors, we can detect anomalies and predict object behaviors. When a detected behavior does not match the learned patterns, it is classed as an anomaly. We can predict an object behavior by matching the observed subbehavior of the object with the learned patterns. Generally, patterns of behaviors in a scene can be constructed by supervised or unsupervised learning of each object's velocities and trajectories, etc. Supervised learning is used for known scenes where objects move in pre-defined ways. For unknown scenes, patterns of behaviors should be constructed by self-organizing and self-learning of image sequences. Fernyhough *et al.* [5] establish the spatio-temporal region by learning results of tracking objects in a image sequence, and construct a qualitative behavior model by qualitative reasoning and statistical analysis.

G. Content-Based Retrieval of Surveillance Videos

The task in content-based retrieval of surveillance videos is to retrieve video clips from surveillance video databases according to video contents, based on automatic image and video understanding. At present, research on video retrieval focuses on the low-level perceptively meaningful representations of pictorial data (such as color, texture, shape, etc) and simple motion information. These retrieval techniques cannot accurately and effectively search the videos for sequences related to specified behaviors. Semantic-based video retrieval (SBVR) aims to bridge the gap between low-level features and high-level semantic meanings. Based on automatic interpretation of contents in surveillance videos, SBVR may classify and further access the surveillance video clips that are related to specific behaviors, and supply a more high-level, more intuitive and more humanistic retrieval mode. Semantic-based retrieval of surveillance videos brings the following difficult problems: automatic extraction of semantic behavior features, combination between low-level visual features and behavior features, hierarchical organization of image and video features, semantic video indexing, inquire interface, etc.

H. Natural Language Description of Object Behaviors

Describing object behaviors by natural language in accord with human habits is a challenging research subject. The key task is to obtain the mapping relationships between object behaviors in image sequences and the natural language. These mapping relationships are related to the following two problems.

1) **Relationships between behaviors and semantic concepts**. Each semantic concept of motion describes a class of behaviors, but each behavior may be related to multiple semantic concepts. After the mapping has been clearly defined, we could construct the relationship between the results of low-level image processing and object behaviors. The key problems include the modeling of semantic concepts of motions, and the automatic learning of semantic concepts of behaviors.
2) **Semantic recognition and natural language description of object behaviors**. People usually describe developments and transformations of objects with concepts at different levels. The higher level concepts require greater background knowledge. It is a key problem to analyze the behaviors of moving objects using the tracking results from low-level systems, and further recognize the more abstract semantic concepts at higher layers. We can use the corresponding relationships between semantic concepts and object behaviors, semantic networks with different layers and reasoning theory to explore this problem. Natural language is the most convenient and natural way for humans to communicate each other. Organizing recognized concepts and further representing object behaviors in brief and clear natural language is one of the ultimate goals of visual surveillance. In addition, the synchronous description, i.e., giving the description before a behavior finishes (during the behavior is progressing), is also a challenge. We should design an incremental description method which is able to predict object behaviors.

I. Fusion of Data From Multiple Sensors

It is obvious that future visual surveillance systems will greatly benefit from the use of multiple cameras [44], [45], [73]. The cooperation between multiple cameras relies greatly on fusion of data from each camera. Data fusion is primarily feature-level based rather than image-level based or decision-making-level based. It happens in single view tracking, correspondence of cross-cameras, automatic camera switching (i.e., best view selection), etc. The main problems involve how to fuse different types of features, e.g., color, geometric features, into one group to track and recognize objects, and further understand their behaviors; how to fuse features extracted from different viewpoints to correspond objects; and how to communicate data about the same object between multiple cameras.

Besides video, sensors for surveillance include audio, infrared, ultrasonic, and radar, etc. Each of these sensors has its own characteristics. Surveillance using multiple different sensors seems to be a very interesting subject. The main problem is how to make use of their respective merits and fuse information from such kinds of sensors.

J. Remote Surveillance

Remote surveillance becomes more and more important for many promising applications, e.g., military combat, prevention of forest fires, etc. Video data are acquired from distributed sensors and transmitted to a remote control center. The transmission process must satisfy the following requirements.

- The upload bandwidth (from sensors to the control center) should be much wider than the download bandwidth (from the control center to sensors).
- The security of transmission must be guaranteed. Because some surveillance data involve privacy, commercial secrets and even national security, and nevertheless are transmitted through public networks, information security becomes a key problem. This needs the developments of the techniques such as digital watermarking and encryption [89].

The demand for remote surveillance and surveillance using multiple cameras and multiple sensors motivates the combination of network and visual surveillance, which brings new challenges in intelligent surveillance.

VIII. Conclusions

Visual surveillance in dynamic scenes is an active and important research area, strongly driven by many potential and promising applications, such as access control in special areas, person-specific identification in certain scenes, crowd flux statistics and congestion analysis, and anomaly detection and alarming, etc.

We have presented an overview of recent developments in visual surveillance within a general processing framework for visual surveillance systems. The state-of-the-art of existing methods in each key issue is described with the focus on the following tasks: detection, tracking, understanding and description of behaviors, personal identification for visual surveillance, and interactive surveillance using multiple cameras. As for the detection of moving objects, it involves environmental modeling, motion segmentation and object classification. Three techniques for motion segmentation are addressed: background subtraction, temporal differencing, and optical flow. We have discussed four intensively studied approaches to tracking: region based, active-contour based, feature based, and model based. We have reviewed several approaches to behavior understanding, including DTW, FSM, HMMs, and TDNN. In addition, we examine the state-of-the-art of behavior description. As to personal identification at a distance, we have divided gait recognition methods into four classes: mode based, statistics, physical-parameter based, and spatio-temporal motion based. As to fusion of data from multiple cameras, we have reviewed installation, object matching, switching, and data fusion.

At the end of this survey, we have given some detailed discussions on future directions, such as occlusion handling, fusion of 2-D tracking and 3-D tracking, 3-D modeling of humans and vehicles, combination of visual surveillance and personal identification, anomaly detection and behavior prediction, content-based retrieval of surveillance videos, natural language description of object behaviors, fusion of data from multiple sensors, and remote surveillance.

Acknowledgment

The authors thank J. Lou, Q. Liu, H. Ning, M. Hu, D. Xie, and G. Xu from the NLPR for their valuable suggestions and assistance in preparing this paper.

References

[1] R. T. Collins, A. J. Lipton, and T. Kanade, "Introduction to the special section on video surveillance," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 745–746, Aug. 2000.

[2] J. Steffens, E. Elagin, and H. Neven, "Person spotter-fast and robust system for human detection, tracking and recognition," in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, 1998, pp. 516–521.

[3] 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, "A system for video surveillance and monitoring," Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep., CMU-RI-TR-00-12, 2000.

[4] I. Haritaoglu, D. Harwood, and L. S. Davis, "W^4: Real-time surveillance of people and their activities," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 809–830, Aug. 2000.

[5] J. Fernyhough, A. G. Cohn, and D. C. Hogg, "Constructing qualitative event models automatically from video input," *Image Vis. Comput.*, vol. 18, no. 9, pp. 81–103, 2000.

[6] A. Baumberg and D. C. Hogg, "Learning deformable models for tracking the human body," in *Motion-Based Recognition*, M. Shah and R. Jain, Eds. Norwell, MA: Kluwer, 1996, pp. 39–60.

[7] H. Z. Ning, L. Wang, W. M. Hu, and T. N. Tan, "Articulated model based people tracking using motion models," in *Proc. Int. Conf. Multi-Model Interfaces*, 2002, pp. 115–120.

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

[9] T. Olson and F. Brill, "Moving object detection and event recognition algorithms for smart cameras," in *Proc. DARPA Image Understanding Workshop*, 1997, pp. 159–175.

[10] A. J. Lipton, H. Fujiyoshi, and R. S. Patil, "Moving target classification and tracking from real-time video," in *Proc. IEEE Workshop Applications of Computer Vision*, 1998, pp. 8–14.

[11] S. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, "Tracking groups of people," *Comput. Vis. Image Understanding*, vol. 80, no. 1, pp. 42–56, 2000.

[12] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, 1999, pp. 246–252.
[13] D. Meyer, J. Denzler, and H. Niemann, "Model based extraction of articulated objects in image sequences for gait analysis," in *Proc. IEEE Int. Conf. Image Processing*, 1998, pp. 78–81.
[14] J. Barron, D. Fleet, and S. Beauchemin, "Performance of optical flow techniques," *Int. J. Comput.Vis.*, vol. 12, no. 1, pp. 42–77, 1994.
[15] N. Friedman and S. Russell, "Image segmentation in video sequences: a probabilistic approach," in *Proc. 13th Conf. Uncertainty in Artificial Intelligence*, 1997, pp. 1–3.
[16] E. Stringa, "Morphological change detection algorithms for surveillance applications," in *Proc. British Machine Vision Conf.*, 2000, pp. 402–412.
[17] Y. Kuno, T. Watanabe, Y. Shimosakoda, and S. Nakagawa, "Automated detection of human for visual surveillance system," in *Proc. Int. Conf. Pattern Recognition*, 1996, pp. 865–869.
[18] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 781–796, Aug. 2000.
[19] A. J. Lipton, "Local application of optic flow to analyze rigid versus nonrigid motion," in *Proc. Int. Conf. Computer Vision Workshop Frame-Rate Vision*, Corfu, Greece, 1999.
[20] C. Stauffer, "Automatic hierarchical classification using time-base co-occurrences," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, 1999, pp. 335–339.
[21] D. Meyer, J. Psl, and H. Niemann, "Gait classification with HMM's for trajectories of body parts extracted by mixture densities," in *Proc. British Machine Vision Conf.*, 1998, pp. 459–468.
[22] J. G. Lou, Q. F. Liu, W. M. Hu, and T. N. Tan, "Semantic interpretation of object activities in a surveillance system," in *Proc. Int. Conf. Pattern Recognition*, 2002, pp. 777–780.
[23] T. Boult, "Frame-rate multi-body tracking for surveillance," in *Proc. DARPA Image Understanding Workshop*, Monterey, CA, Nov. 1998, pp. 305–308.
[24] J. K. Aggarwal, Q. Cai, W. Liao, and B. Sabata, "Non-rigid motion analysis: articulated & elastic motion," *Comput.Vis. Image Understanding*, vol. 70, no. 2, pp. 142–156, 1998.
[25] I. A. Karaulova, P. M. Hall, and A. D. Marshall, "A hierarchical model of dynamics for tracking people with a single video camera," in *Proc. British Machine Vision Conf.*, 2000, pp. 262–352.
[26] S. Ju, M. Black, and Y. Yaccob, "Cardboard people: a parameterized model of articulated image motion," in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, 1996, pp. 38–44.
[27] S. A. Niyogi and E. H. Adelson, "Analyzing and recognizing walking figures in XYT," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1994, pp. 469–474.
[28] K. Rohr, "Toward model-based recognition of human movements in image sequences," *CVGIP: Image Understanding*, vol. 59, no. 1, pp. 94–115, 1994.
[29] S. Wachter and H.-H. Nagel, "Tracking persons in monocular image sequences," *Comput. Vis. Image Understanding*, vol. 74, no. 3, pp. 174–192, 1999.
[30] N. Paragios and R. Deriche, "Geodesic active contours and level sets for the detection and tracking of moving objects," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 266–280, Mar. 2000.
[31] N. Peterfreund, "Robust tracking of position and velocity with Kalman snakes," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 564–569, June 2000.
[32] M. Isard and A. Blake, "Contour tracking by stochastic propagation of conditional density," in *Proc. European Conf. Computer Vision*, 1996, pp. 343–356.
[33] R. Polana and R. Nelson, "Low level recognition of human motion," in *Proc. IEEE Workshop Motion of Non-Rigid and Articulated Objects*, Austin, TX, 1994, pp. 77–82.
[34] D.-S. Jang and H.-I. Choi, "Active models for tracking moving objects," *Pattern Recognit.*, vol. 33, no. 7, pp. 1135–1146, 2000.
[35] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1997, pp. 994–999.
[36] M. Izumi and A. Kojima, "Generating natural language description of human behaviors from video images," in *Proc. Int. Conf. Pattern Recognition*, 2000, pp. 728–731.
[37] A. Kojima, T. Tamura, and K. Fukunaga, "Natural language description of human activities from video images based on concept hierarchy of actions," *Int. J. Comput. Vis.*, vol. 50, no. 2, pp. 171–184, 2002.
[38] G. P. Stein, "Tracking from multiple view points: self-calibration of space and time," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. I, 1999, pp. 521–527.
[39] L. Lee, R. Romano, and G. Stein, "Monitoring activities from multiple video streams: establishing a common coordinate frame," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 758–767, Aug. 2000.
[40] A. Utsumi, H. Mori, J. Ohya, and M. Yachida, "Multiple-view-based tracking of multiple humans," in *Proc. Int. Conf. Pattern Recognition*, 1998, pp. 197–601.
[41] Q. Cai and J. K. Aggarwal, "Tracking human motion using multiple cameras," in *Proc. Int. Conf. Pattern Recognition*, Vienna, Austria, 1996, pp. 68–72.
[42] ——, "Tracking human motion in structured environments using a distributed-camera system," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, no. 11, pp. 1241–1247, 1999.
[43] V. Kettnaker and R. Zabih, "Bayesian multi-camera surveillance," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1999, pp. 253–259.
[44] T. H. Chang, S. Gong, and E. J. Ong, "Tracking multiple people under occlusion using multiple cameras," in *Proc. British Machine Vision Conf.*, 2000, pp. 566–576.
[45] Y. Caspi and M. Irani, "Spatio-temporal alignment of sequences," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 1409–1424, Nov. 2002.
[46] A. Samal and P. A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: a survey," *Pattern Recognit.*, vol. 25, no. 1, pp. 65–77, 1992.
[47] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proc. IEEE*, vol. 83, pp. 705–741, May 1995.
[48] D. Swets and J. Weng, "Discriminant analysis and eigenspace partition tree for face and object recognition from views," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, 1996, pp. 182–187.
[49] B. Moghaddam, W. Wahid, and A. Pentland, "Beyond eigenfaces: probabilistic matching for face recognition," in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, 1998, pp. 30–35.
[50] G. Guo, S. Li, and K. Chan, "Face recognition by support vector machines," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, 2000, pp. 196–201.
[51] H. Rowley, S. Baluja, and T. Kanade, "Neural network based face detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 23–38, Jan. 1998.
[52] C. Garcia and G. Tziritas, "Face detection using quantified skin color regions merging and wavelet packet analysis," *IEEE Trans. Multimedia*, vol. 1, pp. 264–277, Sept. 1999.
[53] B. Menser and M. Wien, "Segmentation and tracking of facial regions in color image sequences," in *Proc. SPIE Visual Communications and Image Processing*, vol. 4067, Perth, Australia, 2000, pp. 731–740.
[54] A. Saber and A. M. Tekalp, "Frontal-view face detection and facial feature extraction using color, shape and symmetry based cost functions," *Pattern Recognit. Lett.*, vol. 19, no. 8, pp. 669–680, 1998.
[55] G. Xu and T. Sugimoto, "A software-based system for real-time face detection and tracking using pan-tilt-zoom controllable camera," in *Proc. Int. Conf. Pattern Recognition*, 1998, pp. 1194–1197.
[56] C. BenAbdelkader, R. Culter, and L. Davis, "Stride and cadence as a biometric in automatic person identification and verification," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, Washington, DC, 2002, pp. 372–377.
[57] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer, "Multi-camera multi-person tracking for EasyLiving," in *Proc. IEEE Int. Workshop Visual Surveillance*, Dublin, Ireland, July 2000, pp. 3–10.
[58] S. A. Niyogi and E. H. Adelson, "Analyzing gait with spatio-temporal surface," in *Proc. IEEE Workshop Motion of Non-Rigid and Articulated Objects*, 1994, pp. 64–69.
[59] C. BenAbdelkader, R. Cutler, H. Nanda, and L. Davis, "EigenGait: motion-based recognition of people using image self-similarity," in *Proc. Int. Conf. Audio- and Video-Based Biometric Person Authentication*, 2001, pp. 312–317.
[60] L. Lee and W. Grimson, "Gait analysis for recognition and classification," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, Washington, DC, 2002, pp. 155–162.
[61] R. Culter and L. Davis, "Robust real-time periodic motion detection, analysis and applications," *IEEE Trans. Pattern Recognit. Machine Intell.*, vol. 13, pp. 129–155, Feb. 2000.
[62] A. Kale, A. Rajagopalan, N. Cuntoor, and V. Kruger, "Gait-based recognition of humans using continuous HMMs," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, Washington, DC, 2002, pp. 336–341.

[63] D. Makris and T. Ellis, "Path detection in video surveillance," *Image Vis. Comput.*, vol. 20, no. 12, pp. 895–903, 2002.
[64] G. Shakhnarovich, L. Lee, and T. Darrell, "Integrated face and gait recognition from multiple views," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001, pp. (I)439–(I)446.
[65] G. Shakhnarovich and T. Darrell, "On probabilistic combination of face and gait cues for identification," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, Washington, DC, 2002, pp. 176–181.
[66] M. S. Nixon, J. N. Carter, D. Cunado, P. S. Huang, and S. V. Stevenage, "Automatic gait recognition," in *BIOMETRICS Personal Identification in Networked Society*, A. K. Jain, Ed. Norwell, MA: Kluwer, 1999, ch. 11.
[67] N. Spencer and J. Carter, "Viewpoint invariance in automatic gait recognition," in *BMVA (British Machine Vision Association) Symp. Advancing Biometric Techniques at the Royal Statistical Society*, London, U.K., Mar. 6, 2002, pp. 1–6.
[68] S. J. Maybank and T. N. Tan, "Special section on visual surveillance—introduction," *Int. J. Comput. Vis.*, vol. 37, no. 2, pp. 173–174, 2000.
[69] C. Regazzoni and V. Ramesh, "Special issue on video communications, processing, and understanding for third generation surveillance systems," *Proc. IEEE*, vol. 89, pp. 1355–1367, Oct. 2001.
[70] M. Köhle, D. Merkl, and J. Kastner, "Clinical gait analysis by neural networks: Issues and experiences," in *Proc. IEEE Symp. Computer-Based Medical Systems*, 1997, pp. 138–143.
[71] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," *IEEE Trans. Pattern Recognit. Machine Intell.*, vol. 23, pp. 349–361, Apr. 2001.
[72] A. Galata, N. Johnson, and D. Hogg, "Learning variable-length Markov models of behavior," *Comput. Vis. Image Understanding*, vol. 81, no. 3, pp. 398–413, 2001.
[73] L. Z. Manor and M. Irani, "Multi-view constraints on homographies," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 214–223, Feb. 2002.
[74] Y. Wu and T. S. Huang, "A co-inference approach to robust visual tracking," in *Proc. Int. Conf. Computer Vision*, vol. II, 2001, pp. 26–33.
[75] L. Wang, W. Hu, and T. Tan, "Recent developments in human motion analysis," *Pattern Recognit.*, vol. 36, no. 3, pp. 585–601, 2003.
[76] S. E. Kemeny, R. Panicacci, B. Pain, L. Matthies, and E. R. Fossum, "Multi-resolution image sensor," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. Aug., pp. 575–583, 1997.
[77] A. Basu and D. Southwell, "Omni-directional sensors for pipe inspection," in *Proc. IEEE Int. Conf. Systems, Man and Cybernetics*, vol. 4, 1995, pp. 3107–3112.
[78] A. Hilton and P. Fua, "Foreword: modeling people toward vision-based understanding of a person's shape, appearance, and movement," *Comput. Vis. Image Understanding*, vol. 81, no. 3, pp. 227–230, 2001.
[79] H. Z. Sun, T. Feng, and T. N. Tan, "Robust extraction of moving objects from image sequences," in *Proc. Asian Conf. Computer Vision*, Taiwan, R.O.C., 2000, pp. 961–964.
[80] W. E. L. Grimson, C. Stauffer, R. Romano, and L. Lee, "Using adaptive tracking to classify and monitor activities in a site," in *Proc. IEEE Conf. Computure Vision and Pattern Recognition*, Santa Barbara, CA, 1998, pp. 22–31.
[81] C. Ridder, O. Munkelt, and H. Kirchner, "Adaptive background estimation and foreground detection using Kalman-filtering," in *Proc. Int. Conf. Recent Advances in Mechatronics*, 1995, pp. 193–199.
[82] D. Koller, J. Weber, T. Huang, J. Malik, G. Ogasawara, B. Rao, and S. Russel, "Toward robust automatic traffic scene analysis in real-time," in *Proc. Int. Conf. Pattern Recognition*, Israel, 1994, pp. 126–131.
[83] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: principles and practice of background maintenance," in *Proc. Int. Conf. Computer Vision*, 1999, pp. 255–261.
[84] H.-Y. Shum, M. Han, and R. Szeliski, "Interactive construction of 3D models from panoramic mosaics," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Santa Barbara, CA, 1998, pp. 427–433.
[85] T. Tian and C. Tomasi, "Comparison of approaches to egomotion computation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1996, pp. 315–320.
[86] Z. Y. Zhang, "Modeling geometric structure and illumination variation of a scene from real images," in *Proc. Int. Conf. Computer Vision*, Bombay, India, 1998, pp. 4–7.
[87] T. N. Tan, G. D. Sullivan, and K. D. Baker, "Model-based localization and recognition of road vehicles," *Int. J. Comput. Vis.*, vol. 29, no. 1, pp. 22–25, 1998.
[88] ——, "Recognizing objects on the ground-plane," *Image Vis. Comput.*, vol. 12, no. 3, pp. 164–172, 1994.
[89] F. Bartoleni, A. Tefas, M. Barni, and I. Pitas, "Image authentication techniques for surveillance applications," *Proc. IEEE*, vol. 89, pp. 1403–1417, Oct. 2001.
[90] K. Karmann and A. Brandt, "Moving object recognition using an adaptive background memory," in *Time-Varying Image Processing and Moving Object Recognition*, V. Cappellini, Ed. Amsterdam, The Netherlands: Elsevier, 1990, vol. 2.
[91] M. Kilger, "A shadow handler in a video-based real-time traffic monitoring system," in *Proc. IEEE Workshop Applications of Computer Vision*, Palm Springs, CA, 1992, pp. 11–18.
[92] JPL, "Traffic surveillance and detection technology development," Sensor Development Final Rep., Jet Propulsion Laboratory Publication no. 97-10, 1997.
[93] J. Malik, S. Russell, J. Weber, T. Huang, and D. Koller, "A machine vision based surveillance system for Californaia roads," Univ. of California, PATH project MOU-83 Final Rep., Nov. 1994.
[94] J. Malik and S. Russell, "Traffic Surveillance and Detection Technology Development: New Traffic Sensor Technology," Univ. of California, Berkeley, California PATH Research Final Rep., UCB-ITS-PRR-97-6, 1997.
[95] W. F. Gardner and D. T. Lawton, "Interactive model-based vehicle tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, pp. 1115–1121, Nov. 1996.
[96] Z. Q. Liu, L. T. Bruton, J. C. Bezdek, J. M. Keller, S. Dance, N. R. Bartley, and C. Zhang, "Dynamic image sequence analysis using fuzzy measures," *IEEE Trans. Syst., Man, Cybern. B*, vol. 31, pp. 557–571, Aug. 2001.
[97] T. J. Fan, G. Medioni, and G. Nevatia, "Recognizing 3-D objects using surface descriptions," *IEEE Trans. Pattern Recognit. Machine Intell.*, vol. 11, pp. 1140–1157, Nov. 1989.
[98] B. Coifman, D. Beymer, P. McLauchlan, and J. Malik, "A real-time computer vision system for vehicle tracking and traffic surveillance," *Transportation Res.: Part C*, vol. 6, no. 4, pp. 271–288, 1998.
[99] J. Malik and S. Russell, "Traffic surveillance and detection technology development (new traffic sensor technology)," Univ. of California, Berkeley, 1996.
[100] C. A. Pau and A. Barber, "Traffic sensor using a color vision method," in *Proc. SPIE—Transportation Sensors and Controls: Collision Avoidance, Traffic Management, and ITS*, vol. 2902, 1996, pp. 156–165.
[101] B. Schiele, "Vodel-free tracking of cars and people based on color regions," in *Proc. IEEE Int. Workshop Performance Evaluation of Tracking and Surveillance*, Grenoble, France, 2000, pp. 61–71.
[102] Q. Delamarre and O. Faugeras, "3D articulated models and multi-view tracking with physical forces," *Comput. Vis. Image Understanding*, vol. 81, no. 3, pp. 328–357, 2001.
[103] ——, "3D articulated models and multi-view tracking with silhouettes," in *Proc. Int. Conf. Computer Vision*, Kerkyra, Greece, 1999, pp. 716–721.
[104] C. Sminchisescu and B. Triggs, "Covariance scaled sampling for monocular 3D body tracking," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Kauai, HI, 2001, pp. I:447–I:454.
[105] R. Plankers and P. Fua, "Articulated soft objects for video-based body modeling," in *Proc. Int. Conf. Computer Vision*, Vancouver, BC, Canada, 2001, pp. 394–401.
[106] T. Zhao, T. S. Wang, and H. Y. Shum, "Learning a highly structured motion model for 3D human tracking," in *Proc. Asian Conf. Computer Vision*, Melbourne, Australia, 2002, pp. 144–149.
[107] J. C. Cheng and J. M. F. Moura, "Capture and representation of human walking in live video sequence," *IEEE Trans. Multimedia*, vol. 1, pp. 144–156, June 1999.
[108] C. Bregler, "Learning and recognizing human dynamics in video sequences," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, 1997, pp. 568–574.
[109] H. Sidenbladh and M. Black, "Stochastic tracking of 3D human figures using 2D image motion," in *Proc. European Conf. Computer Vision*, Dublin, Ireland, 2000, pp. 702–718.
[110] E. Ong and S. Gong, "A dynamic human model using hybrid 2D-3D representation in hierarchical PCA space," in *Proc. British Machine Vision Conf.*, U.K., 1999, pp. 33–42.
[111] H. Yang, J. G. Lou, H. Z. Sun, W. M. Hu, and T. N. Tan, "Efficient and robust vehicle localization," in *Proc. IEEE Int. Conf. Image Processing*, 2001, pp. 355–358.
[112] D. G. Lowe, "Fitting parameterized 3-D models to images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 441–450, May 1991.
[113] J. Phillips, H. Moon, S. Rizvi, and P. Rause, "The FERET evaluation methodology for face recognition algorithms," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 1090–1104, Oct. 2000.

[114] J. Hoshino, H. Saito, and M. Yamamoto, "A match moving technique for merging CG cloth and human video," *J. Visualiz. Comput. Animation*, vol. 12, no. 1, pp. 23–29, 2001.

[115] J. E. Bennett, A. Racine-Poon, and J. C. Wakefield, "MCMC for nonlinear hierarchical models," in *Markov Chain Monte Carlo in Practice*, W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, Eds. London, U.K.: Chapman and Hall, 1996, pp. 339–357.

[116] M. Isard and A. Blake, "CONDENSATION—Conditional density propagation for visual tracking," *Int. J. Comput. Vis.*, vol. 29, no. 1, pp. 5–28, 1998.

[117] ——, "Condensation: unifying low-level and high-level tracking in a stochastic framework," in *Proc. European Conf. Computer Vision*, vol. 1, 1998, pp. 893–909.

[118] H. Kollnig, H. H. Nagel, and M. Otte, "Association of motion verbs with vehicle movements extracted from dense optical flow fields," in *Proc. European Conf. Computer Vision*, 1994, pp. 338–347.

[119] T. N. Tan and K. D. Baker, "Efficient image gradient based vehicle localization," *IEEE Trans. Image Processing*, vol. 9, pp. 1343–1356, Aug. 2000.

[120] T. N. Tan, G. D. Sullivan, and K. D. Baker, "Pose determination and recognition of vehicles in traffic scenes," in *European Conf. Computer Vision—Lecture Notes in Computer Science*, vol. 1, J. O. Eklundh, Ed., Stockholm, Sweden, 1994, pp. 501–506.

[121] ——, "Fast vehicle localization and recognition without line extraction," in *Proc. British Machine Vision Conf.*, 1994, pp. 85–94.

[122] A. E. C. Pece and A. D. Worrall, "Tracking without feature detection," in *Proc. IEEE Int. Workshop Performance Evaluation of Tracking and Surveillance*, Grenoble, France, 2000, pp. 29–37.

[123] ——, "A statistically-based Newton method for pose refinement," *Image Vis. Comput.*, vol. 16, no. 8, pp. 541–544, June 1998.

[124] D. Koller, K. Daniilidis, and H.-H. Nagel, "Model-based object tracking in monocular image sequences of road traffic scenes," *Int. J. Comput. Vis.*, vol. 10, no. 3, pp. 257–281, 1993.

[125] H. Kollnig and H.-H. Nagel, "3D pose estimation by directly matching polyhedral models to gray value gradients," *Int. J. Comput. Vis.*, vol. 23, no. 3, pp. 283–302, 1997.

[126] M. Haag and H.-H. Nagel, "Combination of edge element and optical flow estimates for 3D-model-based vehicle tracking in traffic image sequences," *Int. J. Comput. Vis.*, vol. 35, no. 3, pp. 295–319, 1999.

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

[128] 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.

[129] A. D. Wilson, A. F. Bobick, and J. Cassell, "Temporal classification of natural gesture and application to video coding," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1997, pp. 948–954.

[130] S. G. Gong and H. Buxton, "Editorial: understanding visual behavior," *Image Vis. Comput.*, vol. 20, no. 12, pp. 825–826, 2002.

[131] F. Bremond and G. Medioni, "Scenario recognition in airborne video imagery," in *Proc. Int. Workshop Interpretation of Visual Motion*, 1998, pp. 57–64.

[132] T. Starner, J. Weaver, and A. Pentland, "Real-time American sign language recognition using desk and wearable computer-based video," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 1371–1375, Dec. 1998.

[133] N. M. Oliver, B. Rosario, and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 831–843, Aug. 2000.

[134] M. Brand and V. Kettnaker, "Discovery and segmentation of activities in video," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 844–851, Aug. 2000.

[135] 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.

[136] 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.

[137] Y. A. Ivanov and A. F. Boblic, "Recognition of visual activities and interactions by stochastic parsing," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 852–872, Aug. 2000.

[138] M. Brand, "Understanding manipulation in video," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, 1996, pp. 94–99.

[139] T. Wada and T. Matsuyama, "Multi-object behavior recognition by event driven selective attention method," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 873–887, Aug. 2000.

[140] N. Johnson and D. Hogg, "Learning the distribution of object trajectories for event recognition," *Image Vis. Comput.*, vol. 14, no. 8, pp. 609–615, 1996.

[141] N. Sumpter and A. Bulpitt, "Learning spatio-temporal patterns for predicting object behavior," *Image Vis. Comput.*, vol. 18, no. 9, pp. 697–704, 2000.

[142] J. Owens and A. Hunter, "Application of the self-organizing map to trajectory classification," in *Proc. IEEE Int. Workshop Visual Surveillance*, 2000, pp. 77–83.

[143] G. Herzog and P. Wazinski, "Visual translator: linking perceptions and natural language descriptions," *Artific. Intell. Rev.*, vol. 8, no. 2, pp. 175–187, 1994.

[144] T. Huang, D. Koller, J. Malik, G. Ogasawara, B. Rao, S. Russell, and J. Weber, "Automatic symbolic traffic scene analysis using belief networks," in *Proc. National Conf. Artificial Intelligence*, 1994, pp. 966–972.

[145] P. Remagnino, T. Tan, and K. Baker, "Agent orientated annotation in model based visual surveillance," in *Proc. IEEE Int. Conf. Computer Vision*, 1998, pp. 857–862.

[146] M. Mohnhaupy and B. Neumann, "On the use of motion concepts for top-down control in traffic scene," in *Proc. European Conf. Computer Vision*, 1990, pp. 542–550.

[147] *Proc. AAAI Fall Symp. Computational Models for Integrating Language and Vision*, R. K. Srihari, Ed.. Cambridge, MA, November 1995.

[148] P. Remagnino, T. N. Tan, A. D. Worrall, and K. D. Baker, "Multi-agent visual surveillance of dynamic scenes," *Image Vis. Comput.*, vol. 16, no. 8, pp. 529–532, 1998.

[149] I. Pavlidis, V. Morellas, P. Tsiamyrtzis, and S. Harp, "Urban surveillance system: from the laboratory to the commercial world," *Proc. IEEE*, vol. 89, pp. 1478–1497, Oct. 2001.

[150] O. Javed, S. Khan, Z. Rasheed, and M. Shah, "Camera handoff: tracking in multiple uncalibrated stationary cameras," in *Proc. IEEE Workshop Human Motion (HUMO'00)*, Austin, TX, 2000, pp. 113–118.

[151] S. L. Dockstader and A. M. Tekalp, "Multiple camera tracking of interacting and occluded human motion," *Proc. IEEE*, vol. 89, pp. 1441–1455, Oct. 2001.

[152] R. T. Collins, A. J. Lipton, H. Fujiyoshi, and T. Kanade, "Algorithms for cooperative multi-sensor surveillance," *Proc. IEEE*, vol. 89, pp. 1456–1477, Oct. 2001.

[153] H. Tsutsui, J. Miura, and Y. Shirai, "Optical flow-based person tracking by multiple cameras," in *Proc. IEEE Conf. Multisensor Fusion and Integration in Intelligent Systems*, 2001, pp. 91–96.

[154] A. Mittal and L. S. Davis, "M2 tracker: a multi-view approach to segmenting and tracking people in a cluttered scene," in *Proc. European Conf. Computer Vision*, vol. 1, 2002, pp. 18–36.

[155] J. M. Ferryman, A. D. Worrall, G. D. Sullivan, and K. D. Baker, "A generic deformable model for vehicle recognition," in *Proc. British Machine Vision Conf.*, 1995, pp. 127–136.

[156] D. Cunado, M. S. Nixon, and J. N. Carter, "Using gait as a biometric: via phase-weighted magnitude spectra," in *Proc. Int. Conf. Audio- and Video-Based Biometric Person Authentication*, 1997, pp. 95–102.

[157] ——, "Extracting a human gait model for use as a biometric," in *Proc. Inst. Elect. Eng. (IEE) Colloq. Computer Vision for Virtual Human Modeling*, 1998, pp. 11/1–11/4.

[158] J. M. Nash, J. N. Carter, and M. S. Nixon, "Dynamic feature extraction via the velocity Hough transform," *Pattern Recognit. Lett.*, vol. 18, no. 10, pp. 1035–1047, 1997.

[159] C. Y. Yam, M. S. Nixon, and J. N. Carter, "Extended model-based automatic gait recognition of walking and running," in *Proc. Int. Conf. Audio- and Video-Based Biometric Person Authentication*, 2001, pp. 278–283.

[160] ——, "Gait recognition by walking and running: a model-based approach," in *Proc. Asia Conf. Computer Vision*, Melbourne, Australian, 2002, pp. 1–6.

[161] D. Cunado, J. Nash, M. S. Nixon, and J. N. Carter, "Gait extraction and description by evidence gathering," in *Proc. Int. Conf. Audio- and Video-Based Biometric Person Authentication*, 1999, pp. 43–48.

[162] R. Tanawongsuwan and A. Bobick, "Gait recognition from time-normalized joint-angle trajectories in the walking plane," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001, pp. (II)726–(II)731.

[163] H. Murase and R. Sakai, "Moving object recognition in eigenspace representation: gait analysis and lip reading," *Pattern Recognit. Lett.*, vol. 17, no. 2, pp. 155–162, 1996.

[164] P. S. Huang, C. J. Harris, and M. S. Nixon, "Human gait recognition in canonical space using temporal templates," *Proc. Inst. Elect. Eng. (IEE) Vision Image and Signal Processing*, vol. 146, no. 2, pp. 93–100, 1999.

[165] ——, "Comparing different template features for recognizing people by their gait," in *Proc. British Machine Vision Conf.*, 1998, pp. 639–643.
[166] ——, "Canonical space representation for recognizing humans by gait or face," in *Proc. IEEE Southwest Symp. Image Analysis and Interpretation*, 1998, pp. 180–185.
[167] J. D. Shutler, M. S. Nixon, and C. J. Harris, "Statistical gait recognition via temporal moments," in *Proc. IEEE Southwest Symp. Image Analysis and Interpretation*, 2000, pp. 291–295.
[168] L. Lee, "Gait Dynamics for Recognition and Classification," MIT AI Lab, Cambridge, MA, Tech. Rep. AIM-2001-019, 2001.
[169] O. Javed and M. Shah, "Tracking and object classification for automated surveillance," in *Proc. European Conf. Computer Vision*, vol. 4, 2002, pp. 343–357.
[170] A. Bobick and A. Johnson, "Gait recognition using static, activity-specific parameters," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001, pp. (I)423–(I)430.
[171] A. Johnson and A. Bobick, "A multi-view method for gait recognition using static body parameters," in *Proc. Int. Conf. Audio- and Video-Based Biometric Person Authentication*, 2001, pp. 301–311.
[172] A. Bobick and A. Johnson. (2001) Expected Confusion as a Method of Evaluating Recognition Techniques. [Online]GVU Tech. Rep. GIT-GVU-01-10
[173] C. BenAbdelkader, R. Culter, and L. Davis, "Person identification using automatic height and stride estimation," in *Proc. Int. Conf. Pattern Recognition*, vol. 4, Québec, Canada, 2002, pp. 377–380.
[174] J. G. Lou, H. Yang, W. M. Hu, and T. N. Tan, "Visual vehicle tracking using an improved EKF," in *Proc. Asian Conf. Computer Vision*, 2002, pp. 296–301.
[175] W. M. Hu, D. Xie, and T. N. Tan, "A hierarchical self-organizing approach for learning the patterns of motion trajectories," *Chin. J. Comput.*, vol. 26, no. 4, pp. 417–426, 2003.
[176] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Recognit. Machine Intell.*, vol. 23, pp. 257–267, Mar. 2001.
[177] C. BenAbdelkader, R. Culter, and L. Davis, "Motion-based recognition of people in eigengait space," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, Washington, DC, 2002, pp. 267–274.
[178] R. Collins, R. Gross, and J. Shi, "Silhouette-based human identification from body shape and gait," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, Washington, DC, 2002, pp. 366–371.
[179] C. BenAbdelkader and L. Davis, "Detection of load-carrying people for gait and activity recognition," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, Washington, DC, USA, 2002, pp. 378–383.
[180] B. Bhanu and J. Han, "Individual recognition by kniematic-based gait analysis," in *Proc. Int. Conf. Pattern Recognition*, vol. 3, Québec, Canada, 2002, pp. 343–346.
[181] V. Laxmi, J. Carter, and R. Damper, "Biologically-motivated human gait classifiers," in *Proc. IEEE Workshop Automatic Identification Advanced Technologies*, 2002, pp. 17–22.
[182] I. Robledo and S. Sarkar, "Experiments on gait analysis by exploiting nonstationaryity in the distribution of feature relationships," in *Proc. Int. Conf. Pattern Recognition*, vol. 1, Québec, Canada, 2002, pp. 1–4.
[183] C. BenAbdelkader, R. Cutler, and L. Davis, "View-invariant estimation of height and stride for gait recognition," in *Proc. Workshop Biometric Authentication at European Conf. Computer Vision*, 2002, pp. 155–167.
[184] L. Wang, W. M. Hu, and T. N. Tan, "A new attempt to gait-based human identification," in *Proc. Int. Conf. Pattern Recognition*, 2002, pp. 115–118.
[185] L. Wang, H. Z. Ning, and W. M. Hu, "Gait recognition based on procrustes statistical shape analysis," in *Proc. IEEE Int. Conf. Image Processing*, 2002, pp. III/433–III/436.

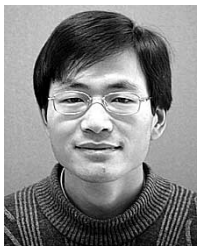
Weiming Hu received the Ph.D. degree from the Department of Computer Science and Engineering, Zhejiang University, Hangzhou, China.

From April 1998 to March 2000, he was a Postdoctoral Research Fellow with the Institute of Computer Science and Technology, Founder Research and Design Center, Peking University, Peking, China. Since April 1998, he has been with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, as an Associate Professor. His research interests are in visual surveillance and monitoring of dynamic scenes, neural networks, and filtering of Internet objectionable images. He has published more than 50 papers in national and international journals and international conferences.

Tieniu Tan (M'92–SM'97–F'03) received the B.Sc. degree in electronic engineering from Xi'an Jiaotong University, China, in 1984 and the M.Sc., DIC, and Ph.D. degrees in electronic engineering from Imperial College of Science, Technology and Medicine, London, U.K., in 1986, 1986, and 1989, respectively.

He joined the Computational Vision Group, Department of Computer Science, The University of Reading, Reading, U.K., in October 1989, where he worked as Research Fellow, Senior Research Fellow, and Lecturer. In January 1998, he returned to China to join the National Laboratory of Pattern Recognition, the Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is currently a Professor and Director of the National Laboratory of Pattern Recognition, as well as President of the Institute of Automation. He has published widely on image processing, computer vision, and pattern recognition. His current research interests include speech and image processing, machine and computer vision, pattern recognition, multimedia, and robotics.

Dr. Tan is an Associate Editor of *Pattern Recognition* and of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and is the Asia Editor of *Image and Vision Computing*. He serves as a referee for many major national and international journals and conferences. He was an elected member of the Executive Committee of the British Machine Vision Association and Society for Pattern Recognition (1996–1997) and is a Founding Co-Chair of the IEEE International Workshop on Visual Surveillance.



Liang Wang received the B.Sc. degree in electrical engineering and the M.Sc. degree in video processing and multimedia communication from Anhui University, Hefei, China, in 1997 and 2000, respectively, and the Ph.D. degree in pattern recognition and intelligent system from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2003.

He has published more than ten papers in major international journals and conferences. His current research interests include computer vision, pattern recognition, digital image processing and analysis, multimedia, and visual surveillance.

Steve Maybank received the B.A. degree in mathematics from King's College, Cambridge, U.K., in 1976 and the Ph.D. degree in computer science from Birkbeck College, University of London, London, U.K., in 1988.

He joined the Pattern Recognition Group, Marconi Command and Control Systems, Frimley, U.K., in 1980 and moved to the GEC Hirst Research Centre, Wembley, U.K., in 1989. During 1993–1995, he was a Royal Society/EPSRC Industrial Fellow in the Department of Engineering Science, University of Oxford, Oxford, U.K. In 1995, he joined the University of Reading, Reading, U.K., as a Lecturer in the Department of Computer Science. In 2004, he became a Professor in the School of Computer Science and Information Systems, Birkbeck College. His research interests include the geometry of multiple images, camera calibration, visual surveillance, information geometry, and the applications of statistics to computer vision.

Dr. Maybank is a Fellow of the Royal Statistical Society and the Institute of Mathematics and its Applications, and is a member of the British Machine Vision Association and the Societe Mathematique de France.