Advances in View-Invariant Human Motion Analysis: A Review

Xiaofei Ji, *Student Member, IEEE*, and Honghai Liu, *Senior Member, IEEE*

***Abstract*—As viewpoint issue is becoming a bottleneck for human motion analysis and its application, in recent years, researchers have been devoted to view-invariant human motion analysis and have achieved inspiring progress. The challenge here is to find a methodology that can recognize human motion patterns to reach increasingly sophisticated levels of human behavior description. This paper provides a comprehensive survey of this significant research with the emphasis on view-invariant representation, and recognition of poses and actions. In order to help readers understand the integrated process of visual analysis of human motion, this paper presents recent development in three major issues involved in a general human motion analysis system, namely, human detection, view-invariant pose representation and estimation, and behavior understanding. Public available standard datasets are recommended. The concluding discussion assesses the progress so far, and outlines some research challenges and future directions, and solution to what is essential to achieve the goals of human motion analysis.**

***Index Terms*—Behavior understanding, human motion analysis, pose representation and estimation, view invariant.**

I. Introduction

HUMAN motion analysis is currently one of the most active research topics in computer vision. This strongly growing interest is driven by a wide spectrum of promising applications in many areas such as visual surveillance, content-based video retrieval, precise analysis of athletic performance, etc., in which the actions are often observed from arbitrary camera viewpoints, for instance, as shown in Fig. 1 [1], and therefore, the present applications request the analysis methods that exhibit some view invariance. This means that analysis methods remain unaffected by different viewpoints of camera. Unfortunately, most of human motion analysis methods are constrained with the assumption of view dependence, i.e., actors have to face a camera or have to be parallel to a viewing plane [2]–[6]. It is evident that such requirements on view dependence are difficult, sometimes impossible, to achieve in realistic scenarios. Research had been conducted to demonstrate the significant role played by the viewpoint in the analysis performance [7], [8]. Due to the limitation of view-dependent characteristic, a large number of human motion analysis methods had been kept away from adapting to a wider application spectrum. It is evident that the viewpoint issue has been one of the bottlenecks for research development and practical implementation of human motion analysis, which has driven growing number of research groups to pay more attentions to research related to the view-invariant issue. A large number of attempts and research progress on removal of the effect on human motion analysis methods had been reported in recent years, especially in view-invariant pose estimation, action representation, and recognition [9]–[11]. Hence, it is timely to comprehensively review recent research on view-invariant human motion analysis.

Fig. 1. Surveillance scene in CMU dataset [1].

Recent research on view-invariant human motion analysis can be characterized by two classes of methods: view-invariant pose representation and estimation, and view-invariant action representation and recognition. The difference is that the former gives priority to the problems of how to estimate 3-D pose from individual image in a sequence, and the latter is focused on the problems of how to infer and understand human action patterns. The two types of methods, however, are closely connected together in that the view-invariant pose estimate usually provides input for the action recognition in order to remove effects caused by view-dependent issues. According to the selection of prior human models, the pose representation and estimation can be categorized into 3-D model-based pose representation and estimation, 3-D model-free pose representation and estimation, and example-based pose representation and estimation [12]–[15]. On the other hand, the research in the view-invariant action representation and recognition can be classified into template-matching-based methods and state-space approaches. The template-based approach is a two-stage method, it first directly investigates view-invariant action

Manuscript received October 15, 2008; revised March 11, 2009. First published August 4, 2009; current version published November 2, 2009. The project is supported in part by the U.K. Engineering and Physical Science Research Council under Grant EP/G041377/1, and in part by the Royal Society under Grant 2008/R1 and Grant 2007/R2. This paper was recommended by Associate Editor E. Trucco.

X. Ji is with the Institute of Industrial Research, The University of Portsmouth, England PO1 3QL, U.K., and also with the College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China. She is also with Shenyang Institute of Aeronautical Engineering, Shenyang 110136, China (e-mail: xiaofei.ji@port.ac.uk).

H. Liu is with the Institute of Industrial Research, the University of Portsmouth, England PO1 3QL, UK (e-mail: honghai.liu@port.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCC.2009.2027608

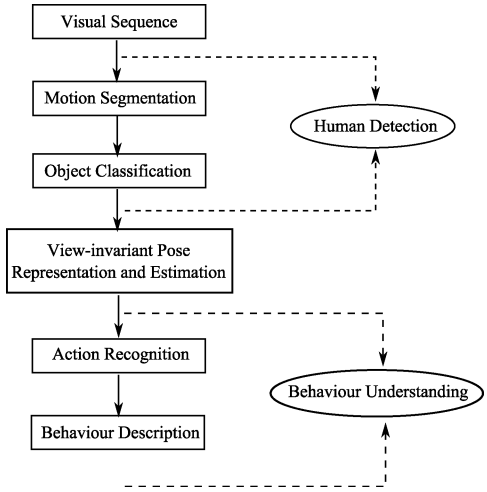

Fig. 2. Framework for view-invariant human motion analysis.

representations, and then considers the action recognition as a classification problem [16], [17]. State-space approaches define each static posture as a state and use certain probabilities to generate mutual connections between these states. Any motion sequence can be considered as a tour though various states of these static postures [18], [19].

It is evident that research on view-invariant human motion analysis has a significant impact on a wide range of applications, particularly, in the following application sectors.

1) Visual Surveillance: To detect, recognize, and track certain objects from image sequences, and to understand and describe object behaviors in dynamic scenarios [20].

2) Content-Based Video Retrieval: To find out where the specified action or event occurs by scanning through a video. Such an application is very useful for sportscasters to quickly retrieve important events in particular games [21].

3) Precise Analysis of Athletic Performance: To automatically analyze complex individual actions of athletes aiming at providing biometric measurements, and visual for coaching assistant and performance improving [22].

The importance and popularity of human motion analysis has led to several previous surveys [23]–[28]. In contrast to the previous reviews, the current review focuses on the most recent developments in human motion analysis, i.e., view-invariant human motion analysis that has not been reviewed at all. In order to provide a comprehensive understanding of visual human motion analysis, a framework is provided, as shown in Fig. 2, which is inspired by research in [25]. It consists of human detection, view-invariant pose representation and estimation, and human behavior understanding. Though priority of this paper is given to recent development in view-invariant human motion analysis, in order to give reader a systematic review, the paper also presents a brief review of human motion detection and behavior description. It is common for an ideal human motion analysis system to have motion detection as the first stage, and high-level behavior description is expected as output. More detailed discussions are provided through this paper on research challenges and future research directions.

The remainder of this paper is organized as follows. Section II reviews research works on human detection including motion segmentation and human body detection. Section III presents research on view-invariant pose presentation and estimation, which is divided into three categories based on selection of a prior human model. Section IV discusses the methods of human behavior understanding. Section V presents selected public available datasets that are dominant for view-invariant research. The paper is concluded in Section VI with analysis of research challenges and directions for future research.

II. Human Detection

Human detection aims at distinguishing moving people from their background. It is a fundamental and crucial issue in a human motion analysis system in that the subsequent processes such as pose estimation and action recognition are greatly dependent on the performance of the human detection [25], [26], [29]. It is impossible to accurately achieve high-level human motion analysis without successful human detection. A brief review is provided as follows in terms of motion segmentation and object classification.

A. Motion Segmentation

Motion segmentation is used to detect regions corresponding to moving objects that can be potential targets in natural scenes in order to provide a focus of attention for later processes such as tracking and activity analyses [26]. Several conventional approaches to motion segmentation are outlined in this section.

1) Background Subtraction: Background subtraction is a widely used approach for motion segmentation, especially under situations with a relatively static background. It attempts to detect moving objects from the difference between the current frame and a reference frame in a pixel-by-pixel fashion. Generally, the reference frame often called the "background image" or "background model" must be a representation of the scene without moving objects, and must be kept regularly updated so as to adapt to the varying luminance conditions and geometry settings [30].

There are several different methods of background subtraction proposed in the recent literature. All of these methods try to effectively estimate the background from the temporal image sequences. The difference in these approaches mainly lies in the type of a background model and the procedure used to update the background model. The simplest background model is based on the temporally averaged image, which is a background approximation to the current static scene. Moreover, Lo and Velastin [31] proposed to use the median value of the last n frames as the background model. This algorithm could handle some of the inconsistencies due to lighting changes.

Recently, some statistical methods have been proposed to extract changing regions from its background. The pioneering work conducted by Stauffer and Grimson [32], who presented an adaptive background mixture model for real-time tracking, in which each pixel was modeled as a mixture of Gaussians and an online approximation, was employed to update the model. The advantages of this kind of methods can be concluded as: first,

it could handle multiple background objects by using multivalued background model, and second, it is insensitive to noise, shadow, and change of lighting conditions. Another challenging problem is how a system consistently focuses on a moving human body and extract it from its background with existence of fast background variations. In order to solve this problem, some researches presented a nonparametric background model that estimates the probability of observing pixel intensity values based on sampling intensity values of individual pixels. This kind of models can deal with situations in which the background of a scene is cluttered and not completely static, but contain small motions from tree branches, bushes, etc. [33]–[35].

Another variant of background subtraction technique is temporal differencing that calculates the pixel difference between two consecutive frames in an image sequence to extract moving regions [36]. It is adaptive to the condition of dynamic environments by making use of previous frame as the current background model. However, temporal differencing works well only if the motion is small. It is common that the methods only detects the outlines of regions of interest, which usually leads to generate holes inside moving entities.

2) Optical Flow: Optical flow is the apparent motion of brightness patterns in the image. Generally, optical flow corresponds to the motion field. Under the assumption of brightness constancy and spatial smoothness, optical flow is used to describe coherent motion of points or features between image frames [37]–[41]. Apart from their vulnerability to image noise, color, and nonuniform lighting, most of flow computation methods have large computational requirements and are sensitive to motion discontinuities.

B. Object Classification

Different moving regions may correspond to different moving objects in natural scenes. Under this assumption, object classification is a necessary process that can analyze moving regions to recognize human being from other moving objects. Roughly speaking, there are two main categories of approaches for classifying moving objects: *shape-based classification* and *motion-based classification*. Shape-based approaches first describe the shape information of moving regions such as points, boxes, blobs, etc. Then it is commonly considered as a standard pattern recognition issue [42]–[47]. However, the articulation of the human body and differences in observed viewpoints lead to a large number of possible appearances of the body, making it difficult to accurately distinguish a moving human from other moving objects using shape-based approach. On the other hand, motion-based approaches directly make use of the periodic property shown in nonrigid articulated human motion to recognize human being from other moving objects. For instance, self-similarity-based time–frequency technology was presented to detect and analyze periodic motion [48]. Furthermore, the shape-based and motion-based approaches were integrated to design a reliable view-independent classification. The results show significant superiority of the hybrid classifier over either motion-based or appearance-based classifiers separately [49], [50]. Fusing multiple features is becoming an important technology to achieve accurate object classification in realistic scenarios.

III. View-Invariant Pose Representation and Estimation

Pose estimation refers to the process of estimating the configuration of the underlying kinematic or skeletal articulation structure of a human. There are a large number of viewpoints from which a human body in a given pose can be observed, each leading to a different appearance of the body. Quantizing the space of viewpoints leads to several view-dependent representations of a single body pose, which causes many limitations during application implementation. View-invariant body-pose representation and estimation is a low-level solution to view-invariant human motion analysis. This section discusses the recent development in detail and classifies them into three categories based on the use of a prior human model.

A. 3-D Model-based Pose Representation and Estimation

3-D model-based pose representation and estimation is the most widely investigated approach to estimate human body pose, and it has progressed significantly in the recent years. Usually, it plays a crucial role in a tracking process in which a model-based analysis-by-synthesis approach is employed. This type of methods uses an explicit 3-D geometric representation of human shape and its kinematic structure to reconstruct a human posture by optimizing the similarity between a model projection and observed images. Generally speaking, 3-D pose estimation in a high-dimensional body configuration space is intrinsically difficult. Hence, the main research is focused on designing search strategies that reduce the solution space.

1) 3-D Human Models: Due to the fact that 2-D human models are restricted to a camera's viewpoint, significant efforts had been paid to depict the geometric structure of human body using 3-D models [12]–[15]. The simplest 3-D representation of a human body is a stick figure, as shown in Fig. 3(a), which consists of line segments linked by joints. This representation is based on the fact that human motion is, essentially, the movement of the supporting bones. This model can then be enhanced with the deformable flesh surrounding the skeletal structure through higher level processing. More complex models include volumetric representations such as generalized cones, ellipses, cylinders, and spheres, as shown in Fig. 3(b). The selection of the models usually depends on the application at hand. It is evident that the more complex the 3-D models, the better the results expected; however, they require more parameters to be estimated, which will lead to more expensive computation cost during the matching process.

2) 3-D Model-Based Pose Estimation: In the recent years, a wide range of estimation techniques for both linear and nonlinear systems had been proposed, and implemented into 3-D model-based pose estimation such as Kalman filter, the condensation algorithm, and dynamic Bayesian network (BN) [53], [54]. Conventionally, Kalman filtering and its variations are proposed due to its efficiency and capability of exact posterior estimation. However, it is a state estimation method based on

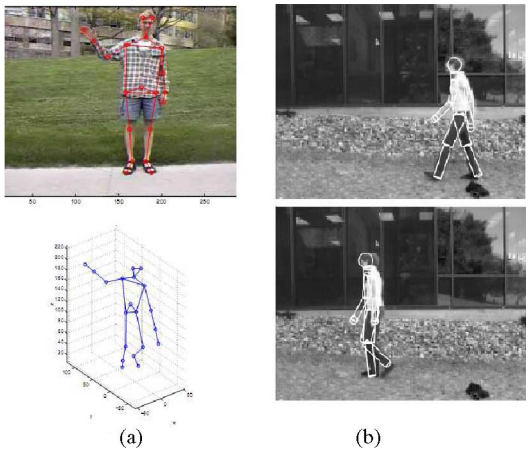

Fig. 3. 3-D human models. (a) Stick figure model [51]. (b) Model with cylinders and spheres [52].

Gaussian distribution, and therefore, it is restricted to situations where the probability distribution of the state parameters is unimodal. In order to cope with clutter situations in which modeling parameters of probability density functions (pdfs) are usually multimodal and non-Gaussian, stochastic sampling strategies are designed to represent simultaneous alternative hypotheses. Among the state-of-the-art in stochastic sampling approaches in visual tracking, the condensation algorithm is the dominant method [53]. It is based upon sampling the posterior distribution estimated in a previous frame, and then, it propagates these samples iteratively to successive images. Recent viewpoint-related research can be categorized into multiple-view 3-D model-based pose estimation and single-view 3-D model-based pose estimation.

Due to the introduction of stochastic sampling and search techniques, the whole-body pose estimation of complex movements can be tracked from multiple views. However, the dimensionality of the state space still remains problematic, and it usually requires a relatively large number of samples to ensure a fair maximum likelihood estimation of a current state. An annealed particle filter was presented by Deutscher *et al.* [55] that combines a deterministic annealing approach with stochastic sampling to reduce the number of samples required. The proposed algorithm was demonstrated in having the capability of efficiently recovering a full articulated body motion. Stochastic metadescent optimization was introduced into a stochastic sampling method for 24-DOF whole-body pose estimation based on multiple views [56]. It can avoid convergence to local minima and achieve efficient performance by using a small number of samples; it is also able to reconstruct complex movements such as kicking and dancing. In addition, Balan *et al.* [57] presented the first quantitative evaluation of Bayesian methods for vision-based 3-D human motion tracking with an emphasis on the effect of prior models, and various likelihood terms in background subtraction and edge measures. A comparison between standard particle filtering and annealed particle filtering was also conducted, which confirmed that both optimization methods worked well for the scenarios consisting of three or more cameras, and also pointed out that the performance of background subtraction dominates the tracking performance. Recently, researchers have been focused on multiview-based real-time tracking and have made some progress in this direction. A representative work was conducted by Caillette *et al.* [13], they first learned Gaussian clusters of given submotions, and then proposed a variable-length Markov model (VLMM) based on the Gaussian clusters that works as an indicator to guide local posture search toward better areas of the distribution. It achieves near real-time performance on long video sequences of ballet dancing scenarios with the advantages of autoinitialization and recovery from tracking failures.

It is evident that human pose reconstruction from a single view image sequence is considerably more difficult than that from multiple views [58]. Besides the difficulty of matching an imperfect, highly flexible, self-occluding model to cluttered image features, realistic body models have at least 30 joint parameters and at least one-third of these DOFs are nearly unobservable in any given monocular image. Sminchisescu *et al.* [59], and Sminchisescu and Triggs [60] had achieved promising results in monocular 3-D human motion capture, which have proved effective on relatively short sequences. Their algorithms represent the probable 3-D configurations of a body over time by propagating a mixture of Gaussians pdf. They make use of performing efficient and thorough global searches on the cost surface associating the image data to potential body configurations. Further, an algorithm for 3-D reconstruction of human action was presented by Loy *et al.* [61] in relatively long monocular image sequences. Such a sequence was represented by a small set of automatically found representative key frames with manually located skeletal joint positions. A 3-D key pose was created for each key frame, and were interpolated among these 3-D body poses incorporating limb length and symmetry constraints, which provided a smooth initial approximation of the 3-D motion. Similarly, detection and tracking techniques were combined to formulate 3-D motion recovery as an interpolation problem from arbitrary viewpoints using a single and potentially moving camera [15]. In addition, a 3-D pose was recovered from single image frames by introducing the use of a probabilistic "proposal map" representing the estimated likelihood of body parts in different 3-D locations with an explicit 3-D model. A data-driven Markov-chain Monte Carlo approach (MCMC) was used to search the high-dimensional pose space. It was applied to estimate 3-D poses of sports players in a variety of complex poses, but it still suffers from high computational cost and difficultly in meeting requirement of reliable analytical detection [62]. It is evident that existing techniques for full-body tracking are far from the use in real-world action recognition. There exist some inherent conditions that need to be taken into account such as occlusions by scene objects, failure recovery, long-term tracking, autoinitialization, generalization to different people, and integration with action recognition. Patrick *et al.* [63] attempted to solve these problems by modeling an action's motions with a variant of the hierarchical hidden Markov model (HMM). Their work made a significant improvement on robustness to observation errors such as occlusion, poor segmentation, and reduced resolution.

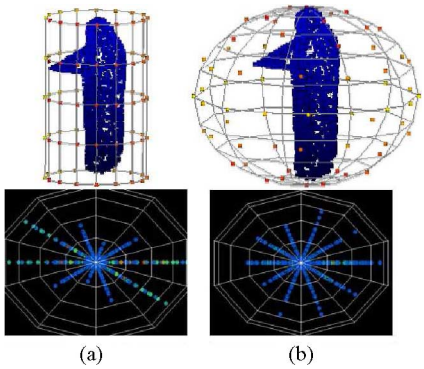

Fig. 4. 3-D visual hull and the shape distribution of the surface using (a) a cylinder and (b) a sphere as reference surface [67].

3-D model-based pose estimation approaches attempt to extract features that explicitly describe position and motion of body parts. These approaches are applicable only in constrained settings, e.g., video productions or sports analysis, but difficult to apply in other scenarios.

B. Model-Free Pose Representation and Estimation

3-D representation is a natural way to deal with the view effect by directly fusing information form multiple images. Such a representation is more informative than simple sets of 2-D images since additional calibration information is required to be taken into account. A number of researchers had investigated 3-D reconstruction of both model shape and motion directly from a visual hull without a prior model [16], [64]–[67]. Visual-hull construction, also known as shape-from-silhouette (SFS), is a popular 3-D reconstruction method that estimates the shape of an object from multiple silhouette images. A representative system was proposed by Mikic *et al.* [64], [65] that integrates automatic acquisition of a human body model with motion tracking from multiple synchronized image sequences. Videos were segmented into individual frames and the 3-D visual-hull reconstruction of a human body shape in each frame was computed in terms of its foreground silhouettes. The resultant reconstruction were further used as input for the model acquisition and tracking algorithms. Additionally, an SFS algorithm was presented to recover shape and joints of a moving articulated object from both its silhouette and color images. Tracking was performed by hierarchically matching the approximate body model to the visual hull using color matching along the silhouette boundary edge [66]. A novel 3-D description of 3-D visual hull was proposed by Cohen and Li [67] for classifying and identifying human posture using a support vector machine. The description is shown in Fig. 4. These kinds of approaches exploit 3-D reconstruction from multiple views to directly recover both shape and motion. Not only is it suitable for estimating complex human movements in a multicamera-based system, but it can also be potentially employed in real-time applications though special hardware is required due to its costly computation.

C. Example-Based Pose Representation and Estimation

Example-based methods are two-stage approaches that first store a database of, for example, human motion figures with known 3-D parameters, and then, estimate a 3-D pose by conducting similarity checking on the examples and an input image [68]–[75]. A potential advantage of example-based methods over model-based method is that a pose can be estimated independently at each frame allowing pose estimation for rapid movements. Shakhnarovich *et al.* [68] presented an example-based approach for view-invariant pose estimation of upper body 3-D poses from a single image. They directly applied parameter-sensitive hashing to rapidly find relevant exemplars of observed image. Experiments demonstrated that the proposed method can rapidly and accurately estimate the articulated poses of human figures from a large database of example images. A method was proposed by Agarwal and Triggs [69], [70] that recovered a 3-D human body pose from monocular silhouettes by direct nonlinear regression of joint angles against histogram-of-shape-context silhouette shape descriptors. Neither a 3-D body model nor labeling of image positions of body parts is required, which makes the method easily adaptable to different people or appearances. Howe [71] proposed a direct silhouette lookup table based on Chamfer distance to select candidate poses that integrate with a Markov chain for temporal propagation for 3-D pose estimation of walking and dancing motions. Another work was proposed by Mori and Malik [72] that match input image with example images to infer the 2-D joint locations, using the technique of shape context matching in conjunction with a kinematic chain-based deformation model. Then the 3-D body configuration and pose are estimated using an existing algorithm. In addition, a learning-based framework was introduced by Elgammal and Lee [73] for inferring 3-D body poses from silhouettes using a single monocular uncalibrated camera. This framework recovered a body pose in a closed form by explicitly learning view-based representations of activity manifolds, and learning mapping functions from such central representation to both the visual input space and the 3-D body pose space, the block diagram for the framework is given in Fig. 5. On the basis of this work, the authors presented a generative model to represent shape deformations according to view and body configuration changes on a 2-D manifold, and then to simultaneously infer 3-D body pose as well as viewpoint from a single silhouette image by learning a nonlinear mapping between manifold embedding and visual input [74]. Regarding real-time issues, a fast pose estimation is achieved by extending gradient boosting techniques to learn a multidimensional map from Haar features to the entire set of joint angles of a full body pose [75].

It is evident that recovering 3-D poses from a single view is a more challenging problem, the aforementioned methods are usually multivalued. Some alternative approaches directly infer a high-level description of the type of pose that the human is performing when it is not necessary to recover 3-D parameters of body joints [9], [18], [19]. Moreover research had been conducted to employ a compact representation of a human action that considered only key poses instead of body poses in all frames. The set of key poses and actions is directly obtained

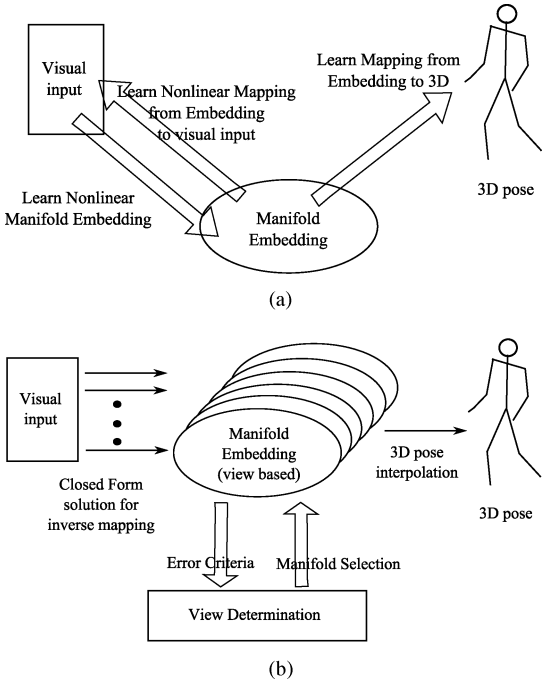

Fig. 5. Proposed learning-based framework in [73]. (a) Learning components. (b) Pose estimation.

from multicamera multiperson training data without manual intervention [18], and silhouette matching between input frames and their key poses is performed using an enhanced pyramid match kernel algorithm (PMK). The PMK can achieve near real-time performance for reasonable coverage of viewpoints.

Example-based approaches that represent the mapping between input images and corresponding pose space provide a powerful mechanism for directly estimating a 3-D pose [27]. The accuracy of the result, however, strongly depends on the similarity of viewpoints for the input images and the example set. Good coverage in a high-dimensional parameter space needs a large number of examples. A current challenge of example-based methods is that these methods were applied in the specific classes of human motions and limited range of viewpoints in training.

Apart from the aforementioned approaches, there are other method based on normalization that remove viewpoint effect by directly transforming all observations into a canonical coordinate frame. These kind of methods must detect the actual motion direction in advance by using the detected body parts or walking direction. Then, matching takes place after the observations have been normalized. For instance, Kale *et al.* [76] proposed a method for view invariant gait recognition, in which a person walk along a straight line (i.e., make a constant angle with the images plane), and then, a side view is synthesized by homography. The viewpoint invariance is also achieved by projecting all the images onto the ground plane [77]. A method was presented by Rogez *et al.* [11] that selects a 2-D viewpoint-insensitive model (made of shape and stick figure), and then uses the 3-D principal directions of man-made environments and the direction of motion to transform both 2-D model and input images to a common frontal view, as shown in Fig. 6. Though these approaches can remove the viewpoint effect directly, a problem with them is that all results completely depend on the robustness of the body orientation estimation. Furthermore, the computational cost is significantly high.

Fig. 6. Original image and transformed image for frontal. (a) Rear-diagonal. (b) Diagonal [11].

IV. Behavior Understanding

Behavior understanding aims to analyze and recognize human motion patterns, to further provide high-level description of human actions and interactions in various scenarios. There are large-scale events that typically depend on the context of environments, objects, or interaction of humans and environments. Generally speaking, behavior understanding can be classified into two categories, which is action recognition and behavior description. Until very recently, most of view-invariant researches are only focused on view-invariant action representation and recognition. However, behavior description is the final goal of many human motion analysis systems, and it is evident that behavior description has been targeted as the most important piece of future work in human motion analysis. Hence, not only does this paper covers the state-of-the-art in view-invariant action recognition, but it also presents recent progress in behavior description.

A. Action Representation and Recognition

Research progresses have been made to attack challenging problems in action representation and recognition of human motion. The challenging problems can be summarized as follows: first, how to deal with the speed of human actions, of which a representation should be independent; second, how to retain the concurrency of the movement of individual body parts, i.e., human body parts move concurrently with possible periodic synchronization. Finally, another challenge is how to overcome variability in actions. For instance, it is evident that a same action executed multiple times by the same person, or by different persons, will exhibit variation in that human actions are not absolutely consistent when they perform a given action [78]. An ideal action representation and recognition system should be able to handle these challenging problems. Most approaches to action representation and recognition are reviewed and discussed in detail in this section.

1) Template-Based Methods: Such an action recognition technique always converts an image sequence into a static shape pattern or a special motion feature pattern, and then, compares it to prestored action prototypes during recognition [23]. The

advantage of template-based methods is low computational cost and simple implementation; however, it is usually more sensitive to noise and variance of movement duration. Bobick and Davis [2] pioneeringly proposed temporal templates to represent human motions, and Hu moments were employed for template matching. An action representation and recognition theory based on motion energy images (MEIs) and motion history images (MHIs) were proposed in their work. Human articulated motions were represented in [3] in terms of spatiotemporal trajectory patterns in individual slices of an image volume XYT, and then, these trajectory patterns are used to classify the motion. Yu *et al.* [4] first directly extracted silhouettes and their contours that are unwrapped and processed by principal component analysis as motion features. Then they employed a three-layer neural network to distinguish the motion patterns into three categories: "walking," "running," and "other" based on the trajectories in eigenspace. In addition, an approach was introduced for measuring the degree of consistency between the implicit underlying motion patterns in two video segments. This was achieved directly from the space-time intensity information in these two video volumes without explicitly computing these motions [79]. Another novel hierarchical model was proposed to classify the different categories of human motion by using a hybrid of spatial–temporal and static features. Experimental results show that compared to previous works, this method offers superior classification performance on a number of large human motion datasets [80]. However, the aforementioned methods and their variants in literature are based on silhouette or contour; therefore, it is to say that these approaches suffer from weaknesses of viewpoint-dependent methodology.

The focus of this section is view-invariant approaches [16], [17], [81]–[85]. Rao *et al.* [81] presented a computational representation of human action to capture dramatic changes in the speed and direction of a motion trajectory, which is presented by spatiotemporal curvature of a 2-D trajectory. This representation is compact, view-invariant, and is capable of explaining an action in terms of meaningful action units. Parameswaran and Chellappa [82], [83] handled the problem of view-invariant action recognition based on point-light displays by investigating geometric invariant theory. These invariants can be computed from five points that lie in a plant, as shown in Fig. 7. They obtained a convenient 2-D invariant representation by decomposing and combining the patches of a 3-D scene. Additionally, a novel action representation was proposed by Yilmaz and Shah [17], [84] by using spatiotemporal action volumes (STVs). Given the object contours of each time slice, an action volume was first generated by computing point correspondences between consecutive contours based on graph theory. Then, a compact action representation in terms of the sign of Gaussian and mean curvatures was obtained by analyzing the differential geometry of the local volume surfaces. Next, the set of these action descriptors was employed to define the action sketch that is invariant to a camera viewing angle. All these methods are based on the assumption that point correspondences are available in parts of images. Therefore, their applications are limited in some special occasions.

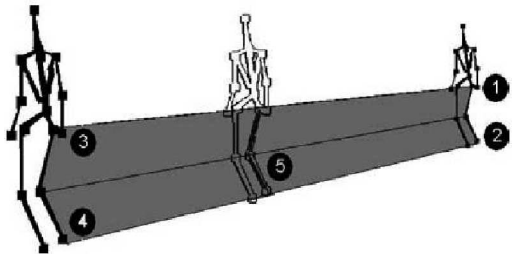

Fig. 7. Geometrical invariants can be computed from five points that lie in a plant [82].

Fig. 8. Space-time shapes of "jumping-jack," "walk," and "run" actions [85].

Another approach was proposed by Blank *et al.* [85] that represented human actions as 3-D shapes induced by the silhouettes in the space-time volume, as shown in Fig. 8. This method extract space-time features such as local space-time saliency, action dynamics, shape structure, and orientation that do not require computing point correspondence. As experiments show, the method is fast and robust to significant changes in scale, partial occlusions, and nonrigid deformations of the actions. This method is not fully view-invariant, it is, however, robust to large changes in viewpoint(up to 54°).

A different approach was presented by Weinland *et al.* [16] that showed fusing multiple view information to achieve view-invariant action recognition. In this paper, first a temporal segmentation method was introduced to split continuous sequence of motions into primitive actions, such as raising and dropping hands and feet, sitting up and down, jumping, etc. [86]. Then, visual hulls were computed and accumulated in a time period between primitive actions into motion history volumes (MHVs), as shown in Fig. 9(a) and (b). Finally, MHVs were transformed into cylindrical coordinates around their vertical axes, and view-invariant features in Fourier space was extracted, as shown in Fig. 9(c) and (d). Results indicated that this representation can be used to learn and recognize basic human action classes, and be independent of gender, body size, and viewpoint.

The key to these template-matching approaches is finding the vital and robust feature sets, and then, an action recognition may be simply considered as a classification problem of these feature sets. All classification algorithms can be applied to human action recognition.

2) State-Space Approaches: Methods based on state-space models usually define each static posture as a state. These states are connected by certain probabilities; any motion sequence is considered as a tour going through various states of these static poses. Probabilities are computed through these tours, and

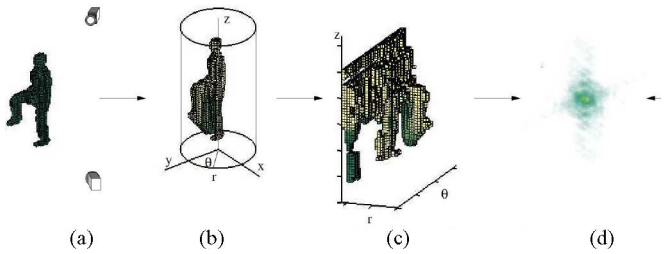

Fig. 9. View-invariant MHV action representation proposed in [16]. (a) Visual hull. (b) MHV. (c) Cylindrical coordinates. (d) Fourier magnitudes.

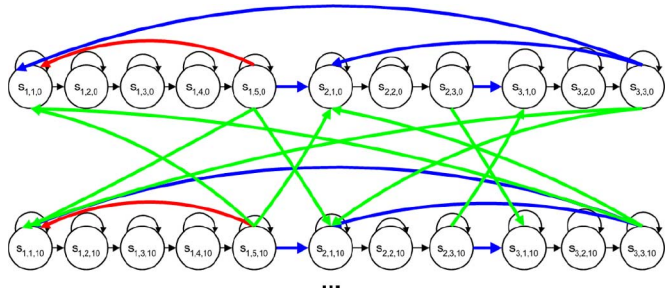

Fig. 10. Action graph models [18].

maximum values are selected as the criteria for human action classification and recognition [23]. State-space methods usually utilize the results of view-invariant pose estimation as input to achieve view-invariant action recognition.

HMMs, a kind of sophisticated techniques for analysing time-varying systems, have been widely applied to express the temporal relationships inherent in human actions [63] [87]–[89]. Lv and Nevatia [90] decomposed a high-dimensional 3-D joint space into a set of feature spaces, in which each feature corresponds to the motion of a single joint or combination of related multiple joints. In the learning process, for each feature, the dynamics of each action class is learned with one HMM. In recognition process, given an image sequence, the observation probability is computed in individual HMMs to recognize each action class, where an AdaBoost scheme is formed to detect and recognize the feature. The proposed algorithm is effective in that the results are convincing in recognizing 22 actions on a large number of motion capture sequences, as well as several annotated and automatically tracked sequences.

Approaches using HMM usually apply intrinsic nonlinear models. It requires searching for a global optimum in the training process, when some conditions are simultaneously considered, such as a large number of actions, viewpoint, etc., which lead to expensive computing iterations. In order to deal with this problem, approaches had been proposed to infer human actions with consideration of contextual constraints imposed by actions. Lv and Nevatia [18] presented an example-based action recognition system that explores the use of contextual constraints. These constraints were inherently modeled by an action representation scheme called *Action Net*, as shown in Fig. 10. In the learning phase, the Action Net is automatically constructed by connection actions with similar boundary key poses. During recognition, silhouette matching between the input frames and the key poses is performed using an enhanced PMK. The best-matched sequence of actions is then tracked using the Viterbi algorithm. This approach was demonstrated on a challenging video sets consisting of 15 complex action classes.

A similar work on example-based HMMs was proposed for view-invariant human motion analysis [19]. This model can account for dependencies between the 3-D exemplars, i.e., representative pose instances and image cues. Inference is then used to identify the action sequence that best explains the image observations. 3-D reconstruction is not required during the recognition phase, and only learned 3-D exemplars are used to produce 2-D image information, which is compared to the observations. This work uses a probabilistic formulation instead of the deterministic-linked action graph introduced in [18], thus allowing to handle uncertainties inherent to actions performed by different people and different styles. The effectiveness of the framework is demonstrated with experiments on real datasets and with challenging recognition scenarios.

HMM and its variants have been widely used to recognition problems such as modeling human motions. However, assumption of independence is usually required in such generative models, which makes the methods unsuitable for accommodating multiple overlapping features or long-range dependencies among observations. Researches have been attempting to introduce conditional random fields (CRFs) to overcome the independence assumption between observations in human motion analysis [91] [92]–[96]. Experimental results show that CRFs can better model dependencies between features and observation than HMMs. On the basic of this research, a novel work was proposed by Natarajan and Nevatia to achieve view and scale invariant action recognition by using multiview shape-flow models, i.e., the shape, flow, and CRF duration [97]. Results demonstrate this method can precisely recognize actions, even with cluttered background. Therefore, it is interesting to investigate how much contribution CRFs and its variants could make to view-invariant human motion analysis.

B. Behavior Semantic Description

The purpose of behavior description is to reasonably choose a group of motion words or shout expressions to report behaviors of moving objects or humans in natural scenes. Generating semantic description is the final goal of human motion analysis; recently, considerable attentions have been paid to this research direction.

Priority has been given to context-free grammar-based approaches. For instance, Ogale *et al.* [9] presented a probabilistic context-free grammar (PCFG) framework to automatically create view-independent representation of actions for multiview training videos. Then, this grammar was used to parse a new single-viewpoint video sequence to deduce the sequence of actions in a view-invariant fashion. Similarly, Yamamoto *et al.* [98] proposed a method to recognize the task-oriented action based on stochastic context-free grammar (SCFG). The proposed method contains many ideas, e.g., recognition of many actions by assigning multiple probabilities, error-free translation for symbolic string, and minimum error classification in sense of Bayesian law.

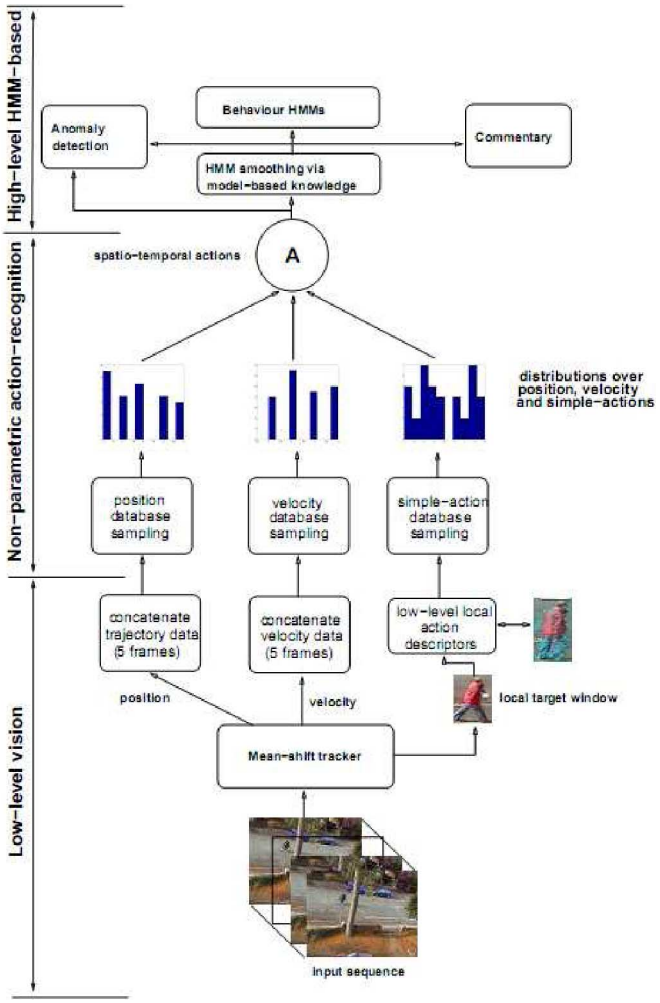

Fig. 11. Diagram of human behavior recognition system [40].

In addition, Park and Aggarwal [99], [100] presented a three-level framework for recognizing human actions and interactions in video. At the low level, the poses of individual body parts including head, torso, arms, and legs are recognized using individual BNs, which are then integrated to obtain an overall body pose. At the middle level, the actions of a single person are modeled in terms of a dynamic BN (DBN). At the high level, the results of midlevel descriptions for each person are juxtaposed along a common time line to identify an interaction between two persons. Spatial and temporal constraints are used for a decision tree to recognize specific interactions. Another system was proposed for human behavior recognition in video sequences. In this system, human actions and behavior are represented using a hierarchy of abstraction: from simple actions, to actions with spatiotemporal context, to action sequences, and finally, to general behaviors. A nonparametric learning and classification technique for actions was combined with an effective parametric representations of action sequence, which was used to describe the behavior. The combined method represents a general framework for human behavior modeling, as shown in Fig. 11. The system demonstrated inspiring results on broadcast tennis sequences for automated video annotation [40].

At present, human behavior description is still restricted to simple and special action patterns, and special scenes. Therefore, research on semantic description of human behaviors in complex unconstrained scenes still remains an open issue.

V. Databases

Standard datasets and protocols are required to fairly evaluate the performance of view-invariant human motion analysis methods. Public available datasets for view-invariant human motion analysis are listed in this section.

1) Carnegie Mellon University (CMU) Motion Database: The CMU motion datasets are constructed by the Robotics Institute of CMU in 2001. The database contains 25 individuals walking on a treadmill in the CMU 3-D room. The subjects perform four different walk patterns consisting of slow walk, fast walk, incline walk, and walk with a ball. All subjects were captured using six high-resolution color cameras distributed evenly around a treadmill [1].

2) Inria Xmas Motion Acquisition Sequence (IXMAS) Dataset: The IXMAS aims to form a dataset comparable to the state-of-the-art in action recognition. It contains 13 actions, each of which was performed three times by 12 actors. In this dataset, the actors freely change their orientation for each acquisition and no further indication on how to perform the actions besides the labels were given [16].

3) HumanEva-I Dataset: It is a human motion dataset for human tracking and pose estimation, and it has been recently made publicly available by the Brown University Group. The dataset contains seven calibrated video sequences (i.e., four of them are grayscale and three are color) that are synchronized with 3-D body poses obtained from a motion capture system. The database contains four subjects performing six common actions (e.g., walking, jogging, gesturing, etc.); for more details, see [101].

4) CASIA Gait Dataset: CASIA is a gait dataset constructed by the Institute of Automation of Chinese Academy of Sciences. It contains three subsets (i.e., datasets A, B, and C). Dataset B is a multiview gait database consisting of 124 subjects and 11 view directions [8]. It can be used to study the relationship between the performance and view angle of human motions, and it also helps to design robust gait recognition systems.

VI. Concluding Remarks

It is evident that view-invariant human motion analysis plays a crucial role in advancing human motion analysis in substantive potential applications such as visual surveillance, content-based video retrieval, precise analysis of athletic performance, etc. This review has presented a comprehensive overview of recent developments in view-invariant human motion analysis with an emphasis on view-invariant pose representation and estimation, and view-invariant action representation and recognition. View-invariant pose representation and estimation is separated into three categories based on the use of a prior human model; view-invariant action representation and recognition is categorized into template-based approaches and state-space approaches. Although a large amount of research has been

conducted in view-invariant human motion analysis, many directions still remain open and problematic, which are outlined as follows.

1) Though significant progresses have been made on model-based pose estimation, research in 3-D pose estimate from monocular sequences still remains problematic, especially the problem of fully automatic initialization of models. Combining example-based and model-based tracking is recommended as a promising direction to solve this problem [15], [102].
2) It is evident that accurately inferring 3-D poses using example-based methods is usually difficult because a large number of parameters needs to be estimated and perspective projection makes recovered poses ambiguous. Existing research have indicated that contextual constraints and multiple-feature fusion methods might provide feasible solutions to the problem [90], [97].
3) Tradeoffs have to be handled between computational cost and recognition accuracy in state-space approaches. New techniques are required to improve their performance and to decrease the computational cost. In comparison with HMMs, CRFs have been advised to view-invariant human motion analysis [97]. Furthermore, state-space approaches should make use of the developments in a 3-D model-based pose estimation to achieve view-invariant human action recognition.
4) Behavior understanding is complex in that a same behavior might have several different meanings depending upon the scene and task context in which it is performed. At present, human behavior description is still restricted to simple and special action patterns and special scenes. Therefore, research on semantic description of human behaviors in complex unconstrained scenes still remains an open issue. Research on behavior patterns constructed by self-organizing and self-learning for unknown scenes is another future research direction.

REFERENCES

[1] R. Gross and J. Shi, "The CMU motion of mody (MoBo) database," Robot. Inst. Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-RI-TR-01-18, Jun. 2001.

[2] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[3] J. Rittscher, A. Blake, and S. Roberts, "Towards the automatic analysis of complex human body motions," *Image Vis. Comput.*, vol. 20, no. 12, pp. 905–916, 2002.

[4] H. Yu, G. Sun, W. Song, and X. Li, "Human motion recognition based on neural network," in *Proc. IEEE Conf. Commun., Circuits Syst.*, 2005, vol. 2, pp. 977–982.

[5] H. Chen, H. Chen, Y. Chen, and S. Lee, "Human action recognition using star skeleton," in *Proc. 4th ACM Int. Workshop Video Surveill. Sens. Netw.*, 2006, pp. 171–178.

[6] H. Li, S. Lin, Y. Zhang, and K. Tao, "Automatic video-based analysis of athlete action," in *Proc. IEEE Conf. Image Anal. Process.*, 2007, pp. 205–210.

[7] O. Masoud and N. Papanikolopoulos, "A method for human action recognition," *Image Vis. Comput.*, vol. 21, no. 8, pp. 729–743, 2003.

[8] S. Yu, D. Tan, and T. Tan, "Modelling the effect of view angle variation on appearance-based gait recognition," in *Proc. Asian Conf. Comput. Vis.*, 2006, vol. 1, pp. 807–816.

[9] A. Ogale, A. Karapurkar, and Y. Aloimonos, "View-invariant modeling and recognition of human actions using grammars," in *Proc. IEEE Conf. Comput. Vis.*, 2005, vol. 5, pp. 115–126.

[10] E. Ong, A. Micilotta, R. Bowden, and A. Hilton, "Viewpoint invariant exemplar-based 3D human tracking," *Comput. Vis. Image Understanding*, vol. 104, no. 2/3, pp. 178–189, 2006.

[11] G. Rogez, J. Guerrero, J. Martinez, and C. Orrite, "Viewpoint independent human motion analysis in man-made environments," in *Proc. Brit. Mach. Vis. Conf.*, 2006, pp. 659–668.

[12] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard, "Tracking loose-limbed people," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, vol. 1, pp. 421–428.

[13] F. Caillette, A. Galata, and T. Howard, "Real-time 3-D human body tracking using variable length Markov models," in *Proc. Brit. Mach. Vis. Conf.*, 2005, pp. 469–478.

[14] C. Menier, E. Boyer, and B. Raffin, "3D skeleton-based body pose recovery," in *Proc. Int. Symp. 3-D Data Process., Vis. Transmiss.*, 2006, pp. 389–396.

[15] A. Fossati, M. Dimitrijevic, V. Lepetit, and P. Fua, "Bridging the gap between detection and tracking for 3D monocular video-based motion capture," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[16] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Comput. Vis. Image Understanding*, vol. 104, pp. 249–257, 2006.

[17] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 984–989.

[18] F. Lv and R. Nevatia, "Single view human action recognition using key pose matching and viterbi path searching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[19] D. Weinland, F. Grenoble, E. Boyer, R. Ronfard, and A. Inc, "Action recognition from arbitrary views using 3D exemplars," in *Proc. IEEE Conf. Comput. Vis.*, 2007, pp. 1–7.

[20] F. Bremond, M. Thonnat, and M. Zuniga, "Video understanding framework for automatic behavior recognition," *Behav. Res. Methods*, vol. 3, no. 38, pp. 416–426, 2006.

[21] Y. Luo, T. Wu, and J. Hwang, "Object-based analysis and interpretation of human motion in sports video sequences by dynamic Bayesian networks," *Comput. Vis. Image Understanding*, vol. 92, no. 2/3, pp. 196–216, 2003.

[22] H. Li, S. Wu, S. Lin, and Y. Zhang, "Automatic detection and recognition of athlete actions in diving video," *Proc. MMM2007*, Singapore, pp. 73–83.

[23] J. Aggarwal and Q. Cai, "Human motion analysis: A review," *Comput. Vis. Image Understanding*, vol. 73, no. 3, pp. 428–440, 1999.

[24] T. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Comput. Vis. Image Understanding*, vol. 81, no. 3, pp. 231–268, 2001.

[25] L. Wang, W. Hu, and T. Tan, "Recent developments in human motion analysis," *Pattern Recognit.*, vol. 36, no. 3, pp. 585–601, 2003.

[26] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. Syst., Man, Cybern. C: Appl. Rev.*, vol. 34, no. 3, pp. 334–352, Aug. 2004.

[27] T. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Understanding*, vol. 104, no. 2/3, pp. 90–126, 2006.

[28] D. Forsyth, O. Arikan, and L. Ikemoto, *Computational Studies of Human Motion: Tracking and Motion Synthesis*. Boston, MA: Now Publishers, 2006.

[29] P. Viola and M. Jones, "Robust real-time object detection," *Int. J. Comput. Vis.*, vol. 2, pp. 882–888, Feb. 2001.

[30] M. Piccardi, "Background subtraction techniques: A review," in *Proc. IEEE Conf. Syst., Man Cybern.*, 2004, vol. 4, pp. 3099–3104.

[31] B. Lo and S. Velastin, "Automatic congestion detection system for underground platforms," in *Proc. Int. Symp. Intell. Multimedia, Video Speech Process.*, 2001, pp. 158–161.

[32] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1999, vol. 2, pp. 246–252.

[33] A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," in *Proc. Eur. Conf. Comput. Vis.—Part II*, 2000, pp. 751–767.

[34] A. Elgammal, R. Duraiswami, and L. Davis, "Efficient kernel density estimation using the fast Gauss transform with applications to color modeling

and tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 11, pp. 1499–1504, Nov. 2003.
[35] B. Han, D. Comaniciu, and L. Davis, "Sequential kernel density approximation through mode propagation: Applications to background modeling," in *Proc. Asian Conf. Comput. Vis.*, 2004, pp. 1–7.
[36] I. Haritaoglu, D. Harwood, and L. Davis, "W (4): Real-time surveillance of people and their activities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 809–830, Aug. 2000.
[37] G. Bradski and J. Davis, "Motion segmentation and pose recognition with motion history gradients," *Mach. Vis. Appl.*, vol. 13, pp. 174–184, 2002.
[38] J. Gonzalez, I. Lim, P. Fua, and D. Thalmann, "Robust tracking and segmentation of human motion in an image sequence," in *Proc. IEEE Conf. Acoust., Speech, Signal Process.*, 2003, vol. 3, pp. 29–32.
[39] A. Efros, A. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. IEEE Conf. Comput. Vis.*, 2003, pp. 726–733.
[40] N. Robertson and I. Reid, "Behaviour understanding in video: A combined method," in *Proc. IEEE Conf. Comput. Vis.*, 2005, vol. 1, pp. 808–815.
[41] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
[42] A. Lipton, H. Fujiyoshi, and R. Patil, "Moving target classification and tracking from real-time video," in *Proc. IEEE Workshop Appl. Comput. Vis.*, 1998, pp. 8–14.
[43] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 4, pp. 349–361, Apr. 2001.
[44] Q. Zhou and J. Aggarwal, "Tracking and classifying moving objects from video," in *Proc. IEEE Workshop Perform. Eval. Tracking Surveill.*, 2001, pp. 1–8.
[45] O. Javed and M. Shah, "Tracking and object classification for automated surveillance," in *Proc. Eur. Conf. Comput. Vis.—Part IV*, 2002, pp. 343–357.
[46] E. Rivlin, M. Rudzsky, R. Goldenberg, U. Bogomolov, and S. Lepchev, "A real-time system for classification of moving objects," in *Proc. IEEE Conf. Pattern Recognit.*, 2002, vol. 3, pp. 688–691.
[47] M. Rodriguez and M. Shah, "Detecting and segmenting humans in crowded scenes," in *Proc. Int. Conf. Multimedia*, 2007, pp. 353–356.
[48] R. Cutler and L. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 781–796, Aug. 2000.
[49] Y. Bogomolov, G. Dror, S. Lapchev, E. Rivlin, M. Rudzsky, and I. Tel-Aviv, "Classification of moving targets based on motion and appearance," in *Proc. Brit. Mach. Vis. Conf.*, 2003, vol. 2, pp. 429–438.
[50] P. Viola, M. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *Int. J. Comput. Vis.*, vol. 63, no. 2, pp. 153–161, 2005.
[51] N. Howe, M. Leventon, and W. Freeman, "Bayesian reconstruction of 3D human motion from single-camera video," *Adv. Neural Inf. Process. Syst.*, vol. 12, pp. 1–8, 1999.
[52] H. Sidenbladh, M. Black, and D. Fleet, "Stochastic tracking of 3D human figures using 2D image motion," in *Proc. Eur. Conf. Comput. Vis.—Part II*, 2000, pp. 702–718.
[53] M. Isard and A. Blake, "Condensation conditional density propagation for visual tracking," *Int. J. Comput. Vis.*, vol. 29, no. 1, pp. 5–28, 1998.
[54] F. Daum, "Nonlinear filters: Beyond the Kalman filter," *IEEE Trans. Aerosp. Electron. Syst. Mag.*, vol. 20, no. 8, pp. 57–69, Aug. 2005.
[55] J. Deutscher, A. Blake, and I. Reid, "Articulated body motion capture by annealed particle filtering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2000, vol. 2, pp. 126–133.
[56] R. Kehl, M. Bray, and L. V. Gool, "Full body tracking from multiple views using stochastic sampling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 129–136.
[57] A. Balan, L. Sigal, and M. Black, "A quantitative evaluation of video-based 3D person tracking," in *Proc. IEEE Workshop Vis. Surveill. Perform. Eval. Tracking Surveill.*, 2005, pp. 349–356.
[58] C. Sminchisescu, "3D human motion analysis in monocular video techniques and challenges," in *Proc. IEEE Conf. Video Signal Based Surveill.*, 2006, pp. 76–100.
[59] C. Sminchisescu, B. Triggs, I. Gravir, and F. Montbonnot, "Covariance scaled sampling for monocular 3D body tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2001, vol. 1, pp. 447–454.
[60] C. Sminchisescu and B. Triggs, "Kinematic jump processes for monocular 3D human tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2003, vol. 1, pp. 69–76.
[61] G. Loy, M. Eriksson, J. Sullivan, and S. Carlsson, "Monocular 3D reconstruction of human motion in long action sequences," in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 442–455.
[62] M. Lee and I. Cohen, "Proposal maps driven MCMC for estimating human body pose in static images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, vol. 2, pp. 334–341.
[63] P. Patrick, W. Svetha, and V. Geoff, "Tracking as recognition for articulated full body human motion anlysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
[64] I. Mikic, M. Trivedi, E. Hunter, and P. Cosman, "Human body model acquisition and motion capture using voxel data," in *Proc. 2nd Int. Workshop Articulated Motion Deformable Objects*, 2002, pp. 104–118.
[65] I. Mikic, M. Trivedi, E. Hunter, and P. Cosman, "Human body model acquisition and tracking using voxel data," *Int. J. Comput. Vis.*, vol. 53, no. 3, pp. 199–223, 2003.
[66] K. Cheung, S. Baker, and T. Kanade, "Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2003, vol. 1, pp. 77–84.
[67] I. Cohen and H. Li, "Inference of human postures by classification of 3D human body shape," in *Proc. IEEE Int. Workshop. Anal. Model. Faces Gestures*, 2003, pp. 74–81.
[68] G. Shakhnarovich, P. Viola, and T. Darrell, "Fast pose estimation with parameter-sensitive hashing," in *Proc. IEEE Conf. Comput. Vis.*, 2003, pp. 750–757.
[69] A. Agarwal and B. Triggs, "3D human pose from silhouettes by relevance vector regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, vol. 2, pp. 882–888.
[70] A. Agarwal and B. Triggs, "Recovering 3D human pose from monocular images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 44–58, Jan. 2006.
[71] N. Howe, "Silhouette lookup for automatic pose tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, pp. 15–22.
[72] G. Mori and J. Malik, "Recovering 3D human body configurations using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1052–1062, Jul. 2006.
[73] A. Elgammal and C. Lee, "Inferring 3D body pose from silhouettes using activity manifold learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, pp. 681–688.
[74] C. Lee and A. Elgammal, "Simultaneous inference of view and body pose using torus manifolds," in *Proc. Int. Conf. Pattern Recognit.*, 2006, vol. 3, pp. 489–494.
[75] A. Bissacco, M. Yang, and S. Soatto, "Fast human pose estimation using appearance and motion via multi-dimensional boosting regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
[76] A. Kale, A. Chowdhury, and R. Chellappa, "Towards a view invariant gait recognition algorithm," in *Proc. IEEE Conf. Adv. Video Signal Based Surveill.*, 2003, pp. 143–150.
[77] P. Ribeiro, J. Santos-Victor, and P. Lisboa, "Human activity recognition from video: Modeling, feature selection and classification architecture," in *Proc. Int. Workshop Human Activity Recognit. Model.*, 2005, pp. 1–10.
[78] V. Parameswaran, "View-invariant in visual human motion analysis," Ph.D. dissertation, Univ. Maryland, College Park, 2004.
[79] E. Shechtman and M. Irani, "Space-time behavior based correlation," in *Proc. IEEE. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 405–412.
[80] J. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
[81] C. Rao, A. Yilmaz, and M. Shah, "View Invariant representation and recognition of actions," *Int. J. Comput. Vis.*, vol. 50, no. 2, pp. 203–226, 2002.
[82] V. Parameswaran and R. Chellappa, "View invariants for human action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2003, vol. 2, pp. 83–101.
[83] V. Parameswaran and R. Chellappa, "View independent human body pose estimation from a single perspective image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, vol. 2, pp. 16–22.
[84] A. Yilmaz and M. Shah, "Actions as objects: A novel action representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 984–989.
[85] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. IEEE Conf. Comput. Vis.*, 2005, vol. 2, pp. 1395–1402.

[86] D. Weinland, R. Ronfard, and E. Boyer, "Automatic discovery of action taxonomies from multiple views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, vol. 2, pp. 1639–1645.
[87] A. Elgammal, V. Shet, Y. Yacoob, and L. Davis, "Learning dynamics for exemplar-based gesture recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2003, vol. 1, pp. 571–578.
[88] T. Mori, Y. Segawa, M. Shimosaka, and T. Sato, "Hierarchical recognition of daily human actions based on continuous hidden Markov Models," in *Proc. IEEE Conf. Automat. Face Gesture Recognit.*, 2004, pp. 779–784.
[89] Y. Shi, A. Bobick, and I. Essa, "Learning temporal sequence model from partially labeled data," in *Proc. 2006 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1631–1638.
[90] F. Lv and R. Nevatia, "Recognition and segmentation of 3-D human action using HMM and multi-class AdaBoost," in *Proc. Eur. Conf. Comput. Vis.*, 2006, vol. 4, pp. 359–372.
[91] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. IEEE Conf. Mach. Learn. Table Contents*, 2001, pp. 282–289.
[92] C. Sminchisescu, A. Kanaujia, and D. Metaxas, "Conditional models for contextual human motion recognition," *Comput. Vis. Image Understanding*, vol. 104, no. 2/3, pp. 210–220, 2006.
[93] S. Wang, A. Quattoni, L. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, vol. 2, pp. 1521–1527.
[94] C. Sutton, A. McCallum, and K. Rohanimanesh, "Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence Data," *J. Mach. Learn. Res.*, vol. 8, pp. 693–723, 2007.
[95] L. Wang and D. Suter, "Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
[96] H. Ning, W. Xu, Y. Gong, and T. Huang, "Latent pose estimator for continuous action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 1–7.
[97] R. Natarajan and P. Nevatia, "View and scale invariant action recognition using multiview shape-flow models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
[98] M. Yamamoto, H. Mitomi, F. Fujiwara, and T. Sato, "Bayesian classification of task-oriented actions based on stochastic context-free grammar," in *Proc. IEEE Conf. Automat. Face Gesture Recognit.*, 2006, pp. 317–323.
[99] S. Park and J. Aggarwal, "Semantic-level understanding of human actions and interactions using event hierarchy," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, pp. 12–20.
[100] S. Park and J. Aggarwal, "A hierarchical Bayesian network for event recognition of human actions and interactions," *Multimedia Syst.*, vol. 10, no. 2, pp. 164–179, 2004.
[101] L. Sigal and M. Black, "HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion," Brown University, Providence, RI, Tech. Rep. CS-06-08, 2006.
[102] C. Curio and M. Giese, "Combining view-based and model-based tracking of articulated human movements," in *Proc. IEEE Workshop Motion Video Comput.*, 2005, vol. 2, pp. 261–268.

Xiaofei Ji (S'09) received the B.Sc. and M.Sc. degrees from Liaoning Shihua University, Fushun, China, in 2000 and 2003, respectively. She is currently working toward the Ph.D. degree in computer vision at the Institute of Industrial Research, University of Portsmouth, Portsmouth, U.K.

Since 2005, she has been a Lecturer in the School of Automation, Shenyang Institute of Aeronautical Engineering, Shenyang, China. She is also with the College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China. Her current research interests include human motion analysis, computer vision, pattern recognition, and vision surveillance.

Honghai Liu (M'02–SM'06) received the Ph.D. degree in robotics from King's College London, London, U.K., in 2003.

Since September 2005, he has been with the University of Portsmouth, Portsmouth, U.K. He was with the University of London, U.K., and the University of Aberdeen, U.K. He also held Industrial Appointments in system integration. He has authored or coauthored more than 100 research papers. His current research interests include computational intelligence methods and applications with a focus on those approaches that could make contributions to the intelligent connection of perception to action. Dr. Liu has received three Best Paper Awards.