Advances in View-Invariant Human Motion Analysis: A Review

Xiaofei Ji  Student Member, IEEE, 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 which can recognize human motion patterns, to reach increasingly sophisticated levels of human behaviour 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 review presents recent development in three major issues involved in a general human motion analysis system, namely human detection, view-invariant pose representation and estimation, behaviour understanding. Public available standard datasets are recommended. The concluding discussion assesses the progress thus far and outlines some research challenges and future directions, and solution to which is essential to achieve the goals of human motion analysis.

Index Terms—Human motion analysis, View-invariant, Pose representation and estimation, Behaviour understanding

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], so the present applications request the analysis methods that exhibit some view invariance. That is to say 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 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 that the significant role that the viewpoint plays 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. A 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 infer and understand human action patterns, while 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 3D model-based pose representation and estimation, 3D 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 representations, 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 through 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 behaviours 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 sports casters 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, that is view-invariant human motion analysis which 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 behaviour understanding. Though priority of this article 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 behaviour description. It is common for an ideal human motion analysis system to have motion detection as the first stage and high-level behaviour description is expected as output. More detailed discussions are provided through this paper on research challenges and future research directions.

![Diagram of framework for view-invariant human motion analysis](image)

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 behaviour understanding. Section V presents selected public available datasets which 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 achieve accurately 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 which can be potential targets in natural scenes in order to provide a focus of attention for later processes such as tracking and activity analysis [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 those 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 in 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 modelled as a mixture of Gaussians and an on-line 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 multi-valued 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 non-parametric background model which 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 containing small motions from tree branches, bushes etc. [33–35].

Another variant of background subtraction technique is temporal differencing which 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 non-uniform 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 which 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 non-rigid 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 approach 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-INVARIENT 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. 3D Model-based Pose Representation and Estimation

3D model-based pose representation and estimation is the most widely investigated approach to estimate human body pose, 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 3D 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, 3D 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) **3D Human Models:** Due to the fact that 2D human models are restricted to a camera’s viewpoint, significant efforts had been paid to depict the geometric structure of human body using 3D models [12–15]. The simplest 3D 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 3D models, the better results may be expected, however they require more parameters to be estimated, which will lead to more expensive computation cost during the matching process.

2) **3D Model-based Pose Estimation:** In the recent years, a wide range of estimation techniques for both linear and non-linear systems had been proposed and implemented into 3D model-based pose estimation such as Kalman filter, the CONDENSATION algorithm, Dynamic Bayesian Network [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 Gaussian distribution, so it is restricted to situations where the probability distribution of the state parameters is unimodal. In order to cope with clutter situations in which
modelling parameters of probability density functions are usually multi-modal 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, then it propagates these samples iteratively to successive images. Recent viewpoint-related research can be categorized into multiple-view 3D model-based pose estimation and single-view 3D 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, 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] which combines a deterministic annealing approach with stochastic sampling to reduce the number of samples required. The proposed algorithm was demonstrated having the capability of recovering a full articulated body motion efficiently. Stochastic meta descent optimization was introduced into a stochastic sampling method for 24 degrees-of-freedom (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 also is 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 3D 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 multi-view based real-time tracking and have made some progress in this direction. A representative work was conducted by Gaillette et al. [13], they first learned Gaussian clusters of given sub-motions, then proposed a variable-length Markov model (VLMM) based on the Gaussian clusters which works as an indicator to guide local posture search towards better areas of the distribution. It achieves near real-time performance on long video sequences of ballet dancing scenarios with the advantages of auto-initialization 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 a third of these degrees of freedom are nearly unobservable in any given monocular image. Sminchisescu and Triggs [59], [60] had achieved promising results in monocular 3D human motion capture, which have proved effective on relatively short sequences. Their algorithms represent the probable 3D 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 3D 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 3D key pose was created for each key-frame and were interpolated among these 3D body poses incorporating limb length and symmetry constraints, which provided a smooth initial approximation of the 3D motion. Similarly, detection and tracking techniques was combined to formulate 3D motion recovery as an interpolation problem from arbitrary viewpoints using a single and potentially moving camera [15]. In addition, a 3D 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 3D locations with an explicit 3D model. A data driven Markov chain Monte Carlo Approach (MCMC) was used to search the high-dimensional pose space. It was applied to estimate 3D poses of sports players in a variety of complex poses, but it still suffers from high computational cost and difficulty meta requirement on 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, auto-initialization, generalization to different people and integration with action recognition. Patrick et al. [63] attempted to solve those problems by modelling an action’s motions with a variant of the hierarchical hidden Markov model. Their work made a significant improvement on robustness to observation errors such as occlusion, poor segmentation and reduced resolution.

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

B. Model Free Pose Representation and Estimation

3D 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 2D images since additional calibration information is required to be taken into account. A number of researchers had investigated 3D 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 3D reconstruction method which estimates the shape of an object from multiple silhouette images. A representative system was proposed by Mimic 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 3D 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, a SFS algorithm was presented to recover shape and joints of a moving articulated object from both its silhouette and colour images. Tracking was performed by hierarchically matching the approximate body model to the visual-hull using colour matching along the silhouette boundary edge [66]. In [67], a novel 3D description of 3D visual-hull was proposed by Cohen and Li for classifying and identifying human posture using a support vector machine. The description is shown in Fig.4. This kind of approaches exploits 3D reconstruction from multiple views to directly recover both shape and motion. Not only is it suitable for estimating complex human movements in a multi-camera based system, but it also can be potentially employed in real-time applications though special hardware is required due to its costly computation.

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

C. Example-Based Pose Representation and Estimation

Example-based methods are two stage approaches which first store a database of example human motion figures with known 3D parameters, then estimate a 3D 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 3D 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 3D human body pose from monocular silhouettes by direct non-linear regression of joint angles against histogram-of-shape-context silhouette shape descriptors. Neither a 3D body model nor labelling of image positions of body parts is required, which makes the method easily adaptable to different people or appearances. Howe et al. [71] proposed a direct silhouette look-up table based on Chamfer distance to select candidate poses which integrate with a Markov chain for temporal propagation for 3D 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 2D joint locations, using the technique of shape context matching in conjunction with a kinematic chain-based deformation model. Then the 3D 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 3D 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 3D 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 two dimensional manifold, then to simultaneously infer 3D body pose as well as viewpoint from a single silhouette image by learning a non-linear mapping between manifold embedding and visual input [74]. Regarding to real-time issues, a fast pose estimation is achieved by extending gradient boosting techniques to learn a multi-dimensional map from Haar features to the entire set of joint angles of a full body pose [75].

It is evident that recovering 3D poses from a single view is a more challenging problem, the above-mentioned methods are usually multi-valued. 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 3D parameters of body joints [9], [18], [19]. Moreover research had been conducted to employ a compact representation of a human action that only considered key poses instead of body poses in all frames. The set of key poses and actions is directly obtained from multi-camera multi-person training data without manual intervention [18], 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 3D pose [27]. The accuracy of the result, however, strongly depends on the similarity of viewpoints for the input images and 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 those methods were applied in the specific classes of human motions and limited range of viewpoints in training.

Apart from the above-mentioned approaches, there is another kind of methods based on normalization, which remove viewpoint effect by directly transforming all observations into a canonical coordinate frame. This kind of methods must be detected the actual motion direction in advance by using the detected body parts or walking direction. Then matching takes place after the observations have been normalised. For instance, Kale et al [76] proposed a method for view invariant gait recognition, in which a person walking along a straight line (i.e., make a constant angle with the images plane), 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 2D viewpoint-insensitive model (made of shape and stick figure), then uses the 3D principal directions of man-made environments and the direction of motion to transform both 2D 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. BEHAVIOUR UNDERSTANDING

Behaviour understanding aims to analyze and recognize human motion patterns, further to 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, behaviour understanding can be classified into two categories, which is action recognition and behaviour description. Until very recently most of view-invariant researches are only focused on view-invariant action representation and recognition. However behaviour description is the final goal of many human motion analysis systems, it is evident that behaviour description has been targeted as most important piece of future work in human motion analysis. Hence, not only does this review cover the state of the art in view-invariant action recognition, but it also presents recent progress in behaviour 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 pre stored 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, Hu moments were
employed for template matching. An action representation and recognition theory based on motion energy images (MEI) and motion history images (MHI) were proposed in their work. In [3], human articulated motions were represented in terms of spatio-temporal trajectory patterns in individual slices of an image volume XYZ; then these trajectory patterns are used to classify the motion. Yu et al. [4] first directly extracted silhouettes and their contours which 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 those two video volumes without explicitly computing those 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 above-mentioned methods and their variants in literature are based on silhouette or contour, it is to say that those 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 spatio-temporal curvature of a 2D 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. Those invariants can be computed from 5 points that lie in a plant, as shown in Fig. 7. They obtained a convenient 2D invariant representation by decomposing and combining the patches of a 3D scene. Additionally, a novel action representation was proposed by Yilmaz and Shah [17], [84] by using spatio-temporal action volumes (STV). Given the object contours of each time slice, an action volume first was 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 analysing the differential geometry of the local volume surfaces. Next the set of these action descriptors was employed to define the action sketch which is invariant to a camera viewing angle. Those methods are all based on the assumption that point correspondences are available in parts of images. So their applications are limited in some special occasions.

Another approach was proposed by Blank and Gorelick [85] that represented human actions as three-dimensional 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 is robust to significant changes in scale, partial occlusions, and non-rigid deformations of the actions. This method is not fully view-invariant, it is however robust to large changes in viewpoint (up to 54 degree).

A different approach was presented by Weinland et al. [16] that fusing multiple view information to achieve view-invariant action recognition. In this work, 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,b). Finally MHVs was transformed into cylindrical coordinates around their vertical axes and view-invariant features in Fourier space was extracted shown in Fig. 9(c,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, then an action recognition may be simply considered as a classification problem of those 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 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.

Hidden Markov Models (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 et al. [90] decomposed a high dimensional 3D 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 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 non-linear 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. Those constraints were inherently modelled 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 Pyramid Match Kernel algorithm. 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 three dimensional exemplars, i.e. representative pose instances and image cues. Inference is then used to identify the action sequence that best explains the image observations. 3D reconstruction is not required during the recognition phase, only learned 3D exemplars are used to produce 2D image information which is compared to the observations. This work uses a probabilistic formulation instead of the deterministic linked action graph introduced in [18], therefore 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 modelling 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 dependences among observations. Researches have been attempting to introduce conditional random fields (CRFs) to overcome the independence assumption between observations in human motion analysis [91–96]. 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, duration-conditional random field [97]. Results demonstrate this method can precisely recognize actions, even with cluttered background. Thereby it is interesting to investigate how much contribution CRFs and its variants could make to view-invariant human motion analysis.

B. Behaviour Semantic Description

The purpose of behaviour description is to reasonably choose a group of motion words or shout expressions to report behaviours 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 multi-view 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.

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 Bayesian networks (BNs), which are then integrated to obtain an overall body pose. At the middle level, the actions of a single person are modelled in terms of a dynamic Bayesian network (DBN). At the high level, the results of mid-level 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 behaviour recognition in video sequences. In this system human actions and behaviour are represented using a hierarchy of abstraction: from simple actions, to actions with spatio-temporal context, to action sequences and finally general behaviours. A non-parametric learning and classification technique for actions was combined with an effective parametric representations of action sequence, which was used to describe behaviour. The combined method represents a general framework for human behaviour modelling, as shown in Fig.11. The system demonstrated inspiring results on broadcast tennis sequences for automated video annotation [40].

At present, human behaviour description is still restricted to simple and special action patterns and special scenes. Therefore, research on semantic description of human behaviours 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) CMU Motion Database: The CMU motion datasets are constructed by the robotics institute of Carnegie Mellon University in 2001. The database contains 25 individuals walking on a treadmill in the CMU 3D 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 colour cameras distributed evenly around a treadmill [1].

2) The IXMAS Dataset: The Inria Xmas Motion Acquisition Sequences (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 3 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 database contains 4 subjects performing 6 common actions (e.g. walking, jogging, gesturing, etc.), for more details please refer to [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, C). Dataset B is a multi-view 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, 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 substantiative 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, 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 3D 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 3D poses using example-based methods usually is difficult due to that 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) Trade-offs 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 3D model-based pose estimation to achieve view-invariant human action recognition.

4) Behaviour understanding is complex in that a same behaviour might have several different meanings depending upon the scene and task context in which it is performed. At present, human behaviour description is still restricted to simple and special action patterns and special scenes. Therefore, research on semantic description of human behaviours in complex unconstrained scenes still remains an open issue. Research on behaviour patterns constructed by self-organizing and self-learning for unknown scenes is another future research direction.

REFERENCES


