CHAPTER 2 LITERATURE REVIEW

To discuss estimation and tracking approaches, various surveys summarized the work depending on their purposes [30, 31, 32, 33, 34, 35]. In this chapter, we briefly review state of the art in 3D human body estimation and tracking limited to the marker-less based system and tracking of large body parts, e.g. trunk, head, arms and legs. It also covers human model representation, image feature extraction, 3D human body motion estimation, 3D human pose inference and quantitative evaluation.

2.1 Human Model Representation

A number of 3D human pose estimation and tracking approaches are based on an explicit human model that is usually represented by a 3D skeleton model. In human model representation, it requires to represent both an articulated skeleton and a shape that is fleshed over the skeleton structure. According to estimation and tracking approach, there are two main human models: kinematic and part-based models.

2.1.1 Kinematic Model

In a 3D kinematic tree, the human body is denoted by a root node, usually the torso, and segments as shown in Figure 2.1 (a). It is represented by $\mathbf{X} = \{x_r, \theta_1, \theta_2, ..., \theta_N\}$, where $x_r \in \mathbb{R}^6$ is the global coordinate and orientation in the world coordinate of the root. $\theta_i \in \mathbb{R}^3$ is the orientation of the i^{th} segment to its corresponding joint. The orientation allows a maximum of three degrees of freedom (DOFs) per joint. In this sense, it leads to a high dimensional state space. Additionally, a shape is introduced to flesh over the articulated skeleton that is often rigid and to relate the 3D natural human to geometric primitives e.g. spheres [36], cylinders [37, 38, 39] or tapered super-quadrics [40, 41, 7]. The kinematic model is usually used in a generative approach (or top-down approach) that the shape is used to match the human model onto the image observation for similarity measurement .

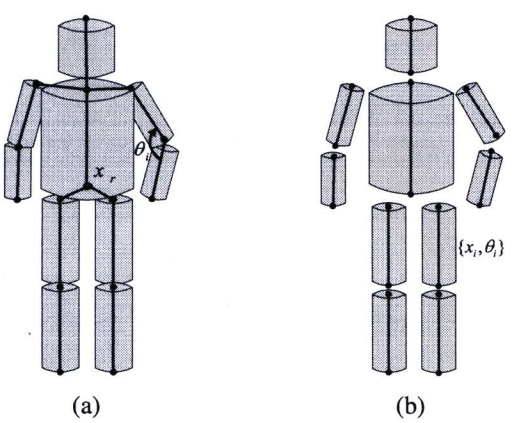

Figure 2.1 Human model representation (a) kinematic model and (b) part-based model

2.1.2 Part-based Model

Apart from a kinematic model, human body can be modeled by a collection of individual parts $\mathbf{X} = \{x_1, \theta_1, x_2, \theta_2, ..., x_N, \theta_N\}$, where x_i and θ_i are position and orientation in the world coordinate of the i^{th} body part, respectively. Various approaches proposed the part-based model, e.g. pictorial structures [42] and loose-limbed model [43]. The shape of part-based model is fleshed by elliptical cones [43] or truncated quadrics [47] as shown in Figure 2.1(b). This model is commonly used in part-based tracking approach that is often based on an inference concept of a graphical model, e.g. belief propagation framework.

2.2 Image Features

Feature is interesting appearance information in an image used as either a mark for tracking or a subject for similarity matching. Such features can be summarized into the following groups:

2.2.1 Edge

An edge feature appears from sharp intensity changes in an image. It can be extracted robustly by a first derivative method and is usually defined as a set of points in the image which has a strong gradient magnitude. It is robust against illumination changing; however, it is still limited to object with texture. A matching function is measured from a normalized distance between edges of human model and the closest edge found in the image [48].

2.2.2 Silhouette

A silhouette feature is a popular feature because it can be simply extracted by background subtraction that attempts to separate moving objects in an image using a reference image by a simple pixel-by-pixel difference concept. Silhouette feature is also robust to cluttered background than the edge feature. Nevertheless, it is limited due to shadow and noisy background segmentation. Moreover, it is difficult to recover 3D human pose using a single camera due to the lack of dept information. A matching function is often based on region overlapping between human model projected onto the silhouette of the input image [49, 50, 48].

2.2.3 Color

A color feature is widely used in human tracking thank to their robustness to rotation especially in depth, shape and scale changing. Several color models are used in tracking applications, e.g. RGB, YUV and HSI color models. However, it usually suffers color changing over time due to changes in scene illumination and visual angle. For robustness against illumination changes, some approaches use hue (H) and saturation (S) components of the HSI space [51]. Lightness (I) is often omitted to make it more robust against global illumination changes. To measure similarity, an appearance of the individual body part is first described using either color histogram [52, 53] or Gaussian color distribution. The similarity measure is defined by the distribution intersection between the color model of the sample and the template. A number of researchers use skin color to extract face and hand regions for initialing state in tracking [52].

2.2.4 Contour

An contour feature refers to the edge representation around a silhouette. It is often referred as active contour that is controlled by an energy function. A matching function is used to minimize the distance between the contour and the edges in the image [54]. A performance of contour is limited by the quality of the silhouette feature, e.g. spurious features due to shadows and noisy background. In 3D human pose inference, a number of researchers employ the contour feature to learn a direct mapping from image features to a 3D human pose [11, 15, 12, 55, 68, 19].

2.2.5 Optical Flow

Optical flow corresponds to the motion information between two consecutive frames. It is formed as a vector field for every pixels in the current frame that independently move from another location in the previous frame. To obtain robust and accurate optical flow feature, it is normally quite computationally expensive, hence it is less frequently used in human motion and pose estimation applications. Bregler [77] represents each pixel by its optical flow and then grouped into blobs with coherent motion. Sminchisescu and Triggs [69] use optical flow to construct an outlier map used to weight the importance of edges.

2.3 3D Human Pose Inference

A number of approaches have been proposed to infer 3D human body pose. According to the number of cameras, these approaches can be divided into two main groups, single view and multi-view groups.

2.3.1 A Single View

In this group, a monocular image is used to reconstruct 3D-articulated human body. It is normally difficult to obtain due to lack of depth information, so various approaches turn to learning concept. Apart from learning-based approaches, 3D human pose inference is recovered by example-based and geometric approaches.

Learning-based Approach Learning-based approach attempts to learn a direct mapping from image features to a 3D human pose [11, 12, 68, 14, 19, 15]. It has learned a probabilistic mapping using training data in the form of Bayesian learning framework [15, 14], regression [55, 68], or Bayesian mixture of experts [12, 16, 17]. The popular features in learning based approach are contour [11, 12, 55, 68, 19] and edge [55] features.

Example-based Approach To avoid training process, some researchers turn to an example-based approach that a set of examples with known poses are stored. The human pose in an input image is estimated by searching for similarity to the database. Several methods are proposed for the searching technique such as parameter sensitive hashing [8], feature matching [9] and database look-up [10]. These methods require extensive database and the performance may be degraded significantly due to difficulties in finding good features.

Geometric Approach This approach attempts to lift 2D joint coresponcing points to

3D model, normally formed as a 3D skeleton model. It assumes that 2D joint points and length of each segment are known a priori. The popular method is usually based on a camera model concept, scaled-orthographic [18, 19, 20, 21, 22, 23] and perspective concepts [25]. In scaled-orthographic concept, it is assumed that most of body parts are close to plane parallel to the image plane. This method is , however, sensitive to perspective effect. Although, performance of the perspective concept is more accurate than that of scaled-orthographic concept, it is limited by an assumption that at least one predefined segment is parallel to the image plane.

2.3.2 Multiple Views

In the multi-view group, images captured from multiple calibrated cameras are used to reconstruct a 3D human pose. It is very helpful for reducing observation ambiguity, for handling self-occlusions and for providing general reliability of data. However, the main drawback of this approach relates to the camera system setup. A visual hull based approach is a popular technique in inferencing 3D human pose using multiple cameras. The main concept is based on a shape-from-silhouette [2, 3, 1, 4, 5, 6, 58]. Each of such different view silhouettes forms, in its projection, a cone, called *visual cone* and an intersection of all these cones forms an approximate 3D bounding geometry of the actual object shape. Once a volumetric representation is created, the 3D human model inference is obtained by fitting human model to the volume. This approach is powerful due to direct availability of 3D information; however, it is very sensitive to noise in silhouettes. To alleviate this problem, Cheung el. al. propose voxel coloring by color consistency across multiple views [58]

Apart from a visual hull based approach, some approaches are based on a top-down concept. They compute the likelihood function of a predefined 3D human model in each image view independently. The product over their likelihoods in each view is taken as an overall similarity measure of the 3D human model [80, 48].

2.4 3D Human Motion Estimation

Traditionally, 3D human motion estimation and tracking approaches can be divided into two main classes: *discriminative and generative classes* [60]. In the discriminative class, these approaches attempt to learn a mapping or searching from image features to 3D human pose either learning-based approach or example-based approach. For the generative class, a human body is typically modeled by a 3D kinematic tree and it consists of two main steps. Pose candidates are first generated and then similarity are measured with an observation image.

According to the direction of estimation, 3D human motion estimation approach can be divided into two main categorizes: top-down and bottom-up approaches [30].

2.4.1 Top-down Approach

The top-down approach is also called the generative approach. Human is normally modeled as a 3D skeleton model and the main concept is an effort to fit a predefined human model to image data. A local search is normally performed around an initial pose estimate [41]. It is usually trapped by local minima due to confusing image

parts in the image . To overcome this problem, various approaches based on multiple hypotheses [62, 44, 39, 45, 64, 65, 66, 48, 46] are introduced. A large number of projected samples are generated and their similarities with respect to the input image are measured and compared. The top-down approaches are generally computationally intensive because of the complex search over the high dimensional state space. The computational cost is associated with an exponential increasing according to the number of body parts, $O(N^m)$ where N is the number of samples for each body part and m is the number of body parts. Moreover, such approaches require a good initialization. To improve effectiveness, various techniques have been proposed.

Dimensionality Reduction Dimensionality reduction is a popular technique to suppress human model representation from high-dimension configuration. The top-down approach with the reduced dimensionality are very efficient because fewer samples are required to adequately approximate the state space posterior distribution. To reduce dimensionality, constraints are usually introduced to restrict the range of human movement e.g. walking parallel to a plane [67]. From the restriction, the joints of shoulders, elbows, hips, knees and ankles are considered with only one DOF. In this sense, the number of configurations is reduced significantly.

Motion Model Prior motion model is commonly employed to reduce the search space by predicting a state and recovering lost tracking. Sidenbladh et al. [39] use trained dynamical models in the prediction step for tracking full body motion. Agarwal and Triggs's method [45] clusters body poses in their training data and then uses them to learn the poses by a local linear autoregression to estimate the human motion. The trained models are generally robust for tracking, but they are restricted for the motions similar to those observed in the training set. Sminchisescu and Triggs [69] introduce Covariance Scaled Sampling (CSS) to guide the particles. Ning et al. [70] use physical motion constraints in the propagation of the particles. Some approaches use a good detection of body part to seed the generative method as an automatic (re)initialization. Lee et al. [50] detect some body parts separately and use them to update subsets of the state space using an analytical inference.

Sample Reduction Various approaches attempt to reduce the number of samples using support vector machines [71], annealed particle filtering [48], hybrid Monte Carlo filtering [72] and kernel particle filtering [73]. Alternatively, a hierarchical search is proposed. The main concept is locating one segment independently and then using its result as a constraint to limit possible solutions of the subsequent joints for the rest in the human model [41, 74].

2.4.2 Bottom-up Approach

In the bottom-up approach, candidate body parts are first extracted from an image and then used to generate possible configurations of human body. The performance of this approach is based on results of feature extraction since it is used to represent body parts. This algorithms do not suffer from (re)initialization problems. A discriminative approach is one of several approaches in the bottom-up group. Apart from discriminative approach, some approaches regarded a human body as a set of blobs [75]. The foreground region is extracted and represented as the body parts. To assemble blob

for human pose representation, it is based on similar characteristics of each blob to the previous frame, e.g. coherent flow [75], color [76], or both [77]. Wren et al. [52] introduce a system, called *Pfinder*, that identifies blobs and refers them as head, hand, torso and leg. They apply two *Pfinder* algorithms to obtain 3D estimates of the hands and head using a human model and kinematic constraints [78]. Sato et al. [79] match blobs of body parts over image sequences using average brightness and rough 3D position obtained by measuring the distance between the center of gravity of the blob and the floor.

To increase performance in assembling each body parts, some approaches employ a graphical model inference concept, called *part-based approach*. The human model is represented by a part-based model, e.g. pictorial structures [42] and loose-limbed model [43]. From the distribution concept, the human model is redundantly represented in a global space leading to complex computation. It is obtained by either a generative method or body part detection. Sigal et al. [80] devise a 3D articulated body tracking by the part-based concept for part detection. They use images from multiple views to facilitate the human pose estimation and tracking problems e.g. occlusion.

2.5 Quantitative Evaluation

The problem of evaluating the estimated pose is limited due to the lack of *ground truth*. In 2D human body tracking, quantitative evaluation is more common and it typically uses hand labeled data [81, 82, 53]. For 3D human pose, it is difficult to obtain the 3D ground truth. Several methods have been proposed for quantitative test. Synthetic data has been extensively used [83, 68, 15, 8, 84]. Another method is applying the estimated motion to a virtual character to show a smooth movement [84].

In this chapter, we present a general survey on related topics in 3D human body estimation and tracking. Detailed review for each topic is described in its corresponding chapter. In the next chapter, we describe mathematical background used in our approach such as graphical model, belief propagation and mode seeking techniques.