CHAPTER 4 OVERVIEW OF THE PROPOSED METHOD

The estimation and tracking of human pose have recently attracted attention from computer vision researchers for several applications such as human-computer interaction, virtual reality, and character animation. These approaches typically focus on the body parts in more detail than tracking human location as a whole. A number of algorithms have been proposed to address human body pose tracking. One of the initial approaches is based on the use of markers that is uncomfortable in many applications due to the use of special equipment which must be installed on the body.

On the other hand, the marker-less system can run without special equipment. Among the marker-less methods, several researchers turn to image-based approaches. In such approaches, features from image(s) served as its input to estimate human pose either 2D or 3D human pose. The 3D human body pose is more realistic than the 2D pose due to lack of depth information in 2D model. Nevertheless, approaches for obtaining 3D human pose are more complex than 2D human body tracking approaches. Various methods are proposed to track the 3D human pose. According to the number of camera views used, these models can be divided into two main groups: multi-view [2, 15, 3, 1, 4, 5, 6, 58] and single view [18, 19, 20, 21, 22, 23, 11, 12, 55, 68, 19, 8, 10, 9]. In the multi-view group, images normally captured from multiple calibrated cameras are used to reconstruct a 3D human pose. Human silhouettes in such different views are used to approximate a volumetric (3D bounding geometry of the actual object shape). Once the volumetric representation is created, the 3D human model inference is obtained by fitting a human model to that volume. The main drawback of this approach relates to camera system setup. Furthermore, some applications do not appropriate for multi-camera system e.g. tracking in movie footage recorded from a single monocular camera.

An alternative approach uses a monocular image to reconstruct a 3D articulated body. Most of single camera approaches are based on training and data set of 3D known human pose. Such approaches require extensive database [8, 9, 10] and training [11, 12, 68, 14, 15, 16, 17]. Moreover, the performance may be degraded significantly due to difficulties in finding good features. Some researchers introduced a generative approach (top-down approach) for 3D human tracking that are generally computationally intensive since they consider all body parts at the same time. Moreover, the generative approach still requires a good initialization.

Some approaches track 2D joint points and then use them to reconstruct 3D human pose by a geometric concept [18, 19, 20, 21, 22, 23, 24, 25]. To recover 3D human pose from 2D point correspondences, it is a difficult problem due to lack of depth information of 2D input data. Most of such approaches use scaled-orthographic concept [18, 19, 20, 21, 22, 23, 24] to reconstruct the 3D human pose. It is a simple method; however, it is restricted to the assumption that most human parts are close to a plane parallel to the image plane in every frame. In [25], they use perspective concept that gives more accurate results than scaled-orthographic concept. However, their approach is still restricted by an assumption that at least one predefined segment lies on a

plane parallel to the image plane. Additionally, these approaches are sensitive to some problems inherit in the monocular approach e.g. self-occlusion and observation ambiguities. Furthermore, all human body tracking approaches either multi-view or single view approach generally face the same problem that is recovering from lost tracking. Details of each problem in the monocular approach are described in Chapter 1.

The challenge to estimate and track 3D human body pose from a monocular image sequence is the motivation and focus of this dissertation. In this dissertation, we propose a marker-less-based approach for 3D articulated relative human estimation and tracking from an uncalibrated monocular image sequence. Our approach is based on reconstruction of 3D relative human pose from 2D point correspondences. The joint points are first extracted from an image input and then used to reconstruct 3D human body pose based on a geometric concept [18, 19, 20, 21, 22, 23, 25]. By this concept, our approach does not require training or data set of known 3D human pose. Moreover, a very accurate solution can always be obtained. This is similar to the inverse kinematics. Additionally, this approach consumes less computational time than the generative approach (top-down approach) of 3D human body tracking [62, 44, 39, 45, 64, 65, 66, 48, 46] because this method considers human body in 2D space that leads to reduced dimensionality of human representation. Although our approach includes inferencing 3D human pose from 2D point correspondences, it takes less computational time thank to our closed-form equations.

This chapter is organized as follows. Section 4.1 begins by a system overview of our approach. The human model in our method is explained in Section 4.2. Section 4.3 and 4.4 describe assumption and initialization system of our approach at the first frame.

4.1 A System Overview

A system overview of our approach is shown in Figure 4.1. In the first frame, an assumption is made for initializing system as shown inside of the upper-right dash box. Like most of the monocular approaches [21, 25, 28, 29], all 2D joint points are assumed to be known a priori in the first frame. Some information are automatically determined for subsequence frames, e.g. color templates of each segment, length of each segment and a reference distance. From the figure, our proposed method consists of two main modules: 2D human body tracking and 3D articulated human reconstruction modules.

In the first module, an image sequence taken by an uncalibrated camera serves as input to estimate and track 2D human body joints. We introduce an efficient part-based approach Quick Shift Belief Propagation (QSBP) to reduce the computational time. Compared with the exponentially increasing number of human model candidates in the generative approach, our method converges with linear complexity of the number of body parts [29, 108, 101, 100, 102]. The generative approach considers all body parts at the same time, while part-based approach considers each body part separately and then combines them into the global solution later.

For the second module, the obtained joint points are then used to reconstruct the 3D

articulated human pose. A set of 3D possible configurations are obtained due to non-uniqueness of solutions. In the next process, an efficient technique, based on an Multiple Hypothesis Tracking (MHT) concept with a motion-smoothness function between consecutive frames, is proposed to find optimal solution from a set of possible solutions. Moreover, we apply feedback information from 3D human pose to alleviate several problems inherit in 2D human body tracking using a single camera in the first module, e.g. self-occlusion and observation ambiguity problems. Additionally, a motion model based on feedback information and geometric constraint is introduced for initializing state and (re)initializing state in case of lost tracking.

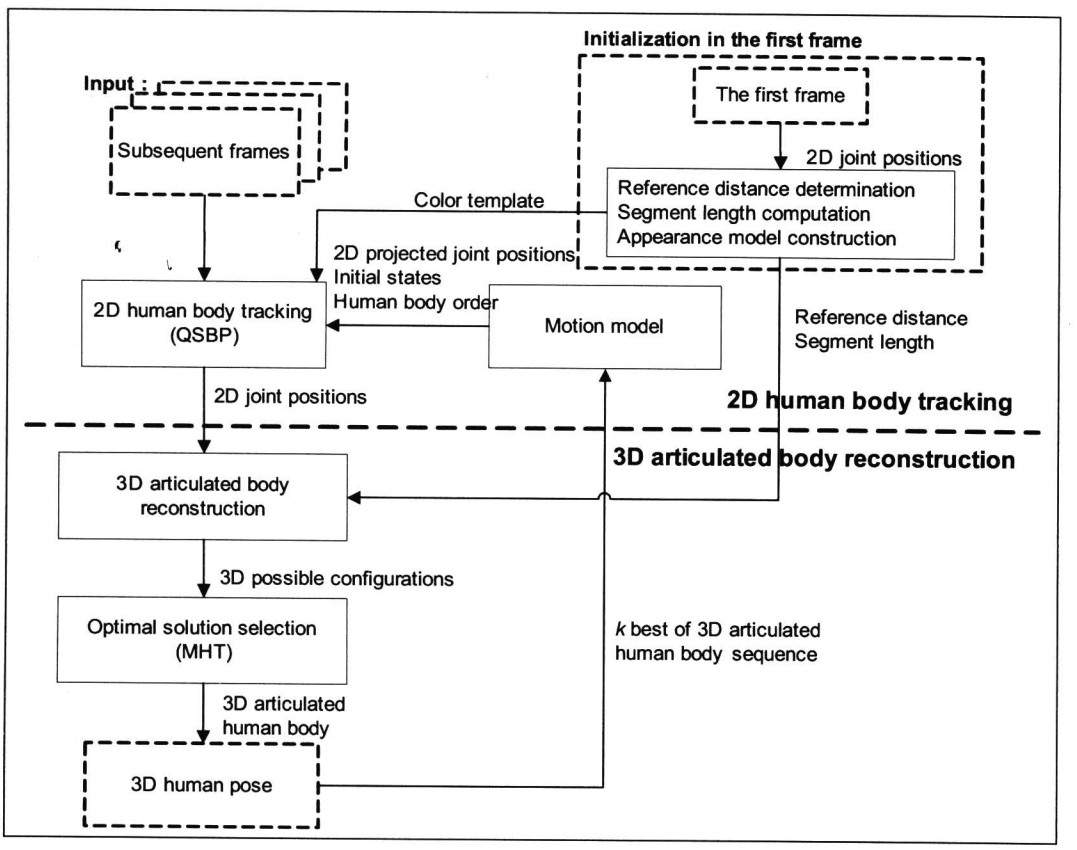

Figure 4.1 Overview of the proposed method.

4.2 Human Model

In articulated human body tracking, the human model is usually represented by an explicit skeleton human model either 2D or 3D model. There are two main human models: kinematic tree and part-based models. In kinematic tree model, human body is denoted by a set of joints and its corresponding segment information. Apart from a kinematic model, the part-based model depicts the human body by a collect of individual parts. Moreover, a shape model is introduced to flesh over the skeleton structure by geometric primitives, e.g. spheres [36], cylinders [37, 38, 39], tapered super-quadrics [40, 41, 7], elliptical cones [43] or truncated quadrics [47].

In our approach, human body is modeled by 3D human skeleton that consists of 15 joint points and 14 segments as shown in Figure 4.2 (a) and (b). The joints are (in

order): abdomen, neck, head, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle and right ankle. Each body part of our human model is fleshed by a planar patch. It is simple shape; however, it can efficiently approximate each rigid human body part.

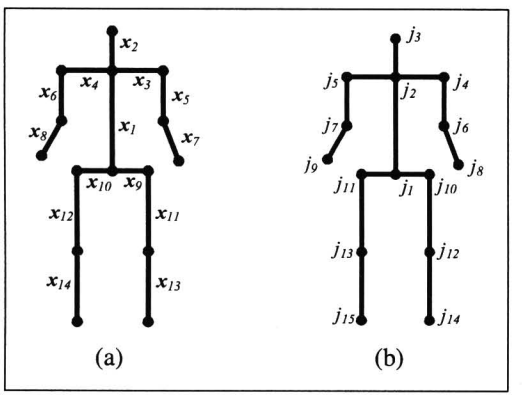

Figure 4.2 The 3D skeleton human model in this dissertation (a) segments (b) joints

4.3 Assumptions

To reconstruct 3D human pose from 2D point correspondences, assumptions or prior knowledge about the human pose or the imaging environment are normally made in most research studies. For example, various assumptions are employed: that most human parts are close to a plane parallel to the image plane [61, 18, 20], that at least one predefined human part is parallel to the image plane [25, 109] or that root joint of human body moves parallel to the image plane without Z direction displacement [110] in every frames. These assumptions are difficult to meet in most general image sequences. Moreover, some approaches assume that camera parameters are known a priori [25].

In our approach, simpler assumptions are only made in the first frame for estimating some parameters in 3D reconstruction, e.g. reference distance and relative length of each body part. We select a set of three prespecified joints forming a right triangle with only one of its legs lain on a plane parallel to the image plane. This assumption is easily achieved compared to the assumption regarding camera parameters, which may or may not be applicable. For example, an actor stands vertically parallel to the image plane and not all of his/her joints lie on a plane parallel to the image plane as shown in Figure 4.3. In our experiments, we select the abdomen, neck and left shoulder joints as points of the right triangle. The abdomen and neck joints of the human body are lain on the plane that is parallel to the image plane; however, not the left shoulder segment.

4.4 Initialization

Like most of the monocular approaches [21, 25, 28, 29], our work requires that all joint points are known a priori in the first frame. From our assumption in the first frame, the templates of each segment are automatically formed by a Gaussian mixture model in HS color space for later use in 2D human body tracking. The relative length of each

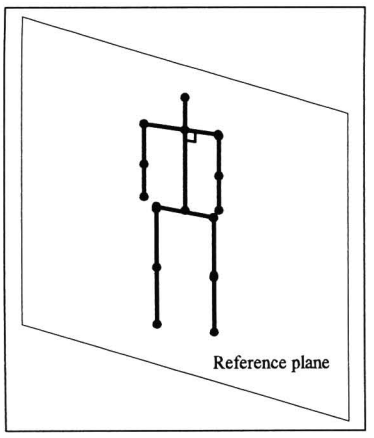

Figure 4.3 Initial posture in the first frame. Note that the segment formed by the neck and the left shoulder joints is not in the reference plane

segment as well as the reference distance are then determined for the reconstruction of a 3D articulated relative human pose. Detailed explanation is described in Section 6.2.

The first module of our approach, 2D human body estimation and tracking, is explained in the Chapter 5. It shows results of 2D human body tracking under self-occlusion and observation ambiguity cases. Moreover, it presents performance of reduced computational time by comparison with other works. Then, the 2D obtained human joint points are used to reconstruct 3D relative human pose that is described in Chapter 6. That chapter shows performance of our approach by evaluation using both synthesized and real-world image sequences. In addition, it also introduces an approach to find the optimal solution to the non-uniqueness problem and also shows robustness and performance comparison with other approaches.