CHAPTER 1 INTRODUCTION

Human body tracking has recently attracted increased attention from computer vision researchers. It is an important subject because it serves as a front-end module for subsequent higher-level processes to understand human activities in several applications, such as human-computer interaction, virtual reality, and character animation. Recently, a number of approaches have been proposed for estimating and tracking 3D human pose either from single or multiple view(s).

Single view approaches usually suffer from self-occlusion, observation ambiguity and non-uniqueness problems due to its fixed viewing angle. Although most of multi-camera-based approaches do not have such limitations, cameras are usually installed at different locations and assumed to be synchronized and calibrated [1, 2, 3, 4, 5, 6]. Such approaches require specialized cameras and expensive hardwares since general cameras do not normally provide these functionalities [7]. The main concept of multi-view-based approaches is merging features in input images from each camera to reconstruct a 3D human pose. Accurate results usually depend on camera setups. Moreover, some applications do not appropriate for multi-camera systems e.g. tracking in movie footage recorded from a single monocular camera. In the monocular approach, it works on an image sequence taken by a general camera that is simpler to obtain. The challenge to accurately and efficiently estimate and track 3D human body pose from a monocular image sequence is an interesting topic to solve and is focused in this dissertation.

1.1 Challenges

The challenges in 3D human pose estimating and tracking from a monocular image sequence are described as follows :

3D Reconstruction Most of 3D reconstruction approaches [8, 9, 10, 11, 12, 68, 14, 15, 16, 17] are based on previously trained 3D human pose that is difficult to obtain and have limited uses. Moreover, such approaches require extensive database and training. The performance may also be degraded significantly due to difficulties in finding good features. Some approaches track 2D joint points and then use them to reconstruct a 3D human pose by some geometric concepts. Figure 1.1 shows an example of the reconstruction. Result of reconstruction of 3D human pose from 2D joint points in the different views is shown in Figure 1.1 (b) - (d). To recover the 3D human pose from 2D point correspondences, it is a difficult problem due to the lack of depth information of 2D input data. In [18, 19, 20, 21, 22, 23, 24], they used scaled-orthographic concept to reconstruct the 3D human pose. Their method is simple; 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 propose a method based on a perspective concept to reconstruct a 3D human pose. This approach is more accurate than that based on the scaled-orthographic concept [18, 19, 20, 21, 22, 23, 24]; however, it is restricted to the assumption that at least one predefined segment is parallel to the image plane in every frame.

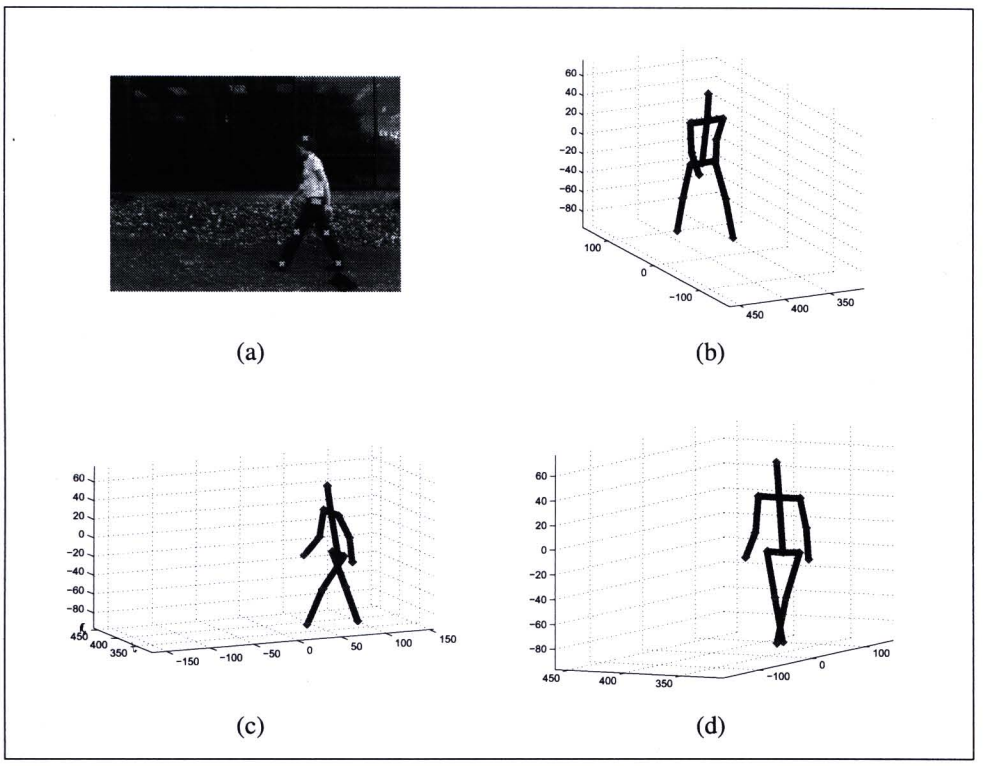

Figure 1.1 Reconstruction of a 3D human pose from 2D point correspondences (a) 2D human joint points are drawn by cross markers and (b) - (d) samples of result from estimation of the 3D human pose from 2D human joint points in the different views

Non-uniqueness Problem One of the main problems in the monocular approach is non-uniqueness of solutions. Generally, a set of 3D possible human poses is obtained using a single camera. It is difficult to select the optimal solution due to its fixed point. Most of monocular approaches are based on choosing the closest configuration from one frame to the next [21, 25]. This method maintains only one solution from the last frame. The tracking may fail due to an incorrect decision in early frames. Moreover, techniques using smooth dynamical model have been introduced to select the optimal solution [26, 27]. However, this method is limited to predefined activities. Figure 1.2 shows an example of the non-uniqueness problem. Figure 1.2 (a) shows an input image with 2D skeleton. Samples of possible 3D human poses due to non-uniqueness problem are shown in Figure 1.2 (b) - (d). In these different 3D poses, each corresponding 3D human joint projects on the same point in the input image.

High-Dimensional Pose Space 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. In the generative approach, the human body is normally represented by a 3D kinematic tree. It is denoted by a set of root, usually the torso, and segment information. The root segment is represented by a global coordinate and orientation in the world coordinate. Each of the other segments is represented by the its orientation from its parent joint to its end joint. The orientation

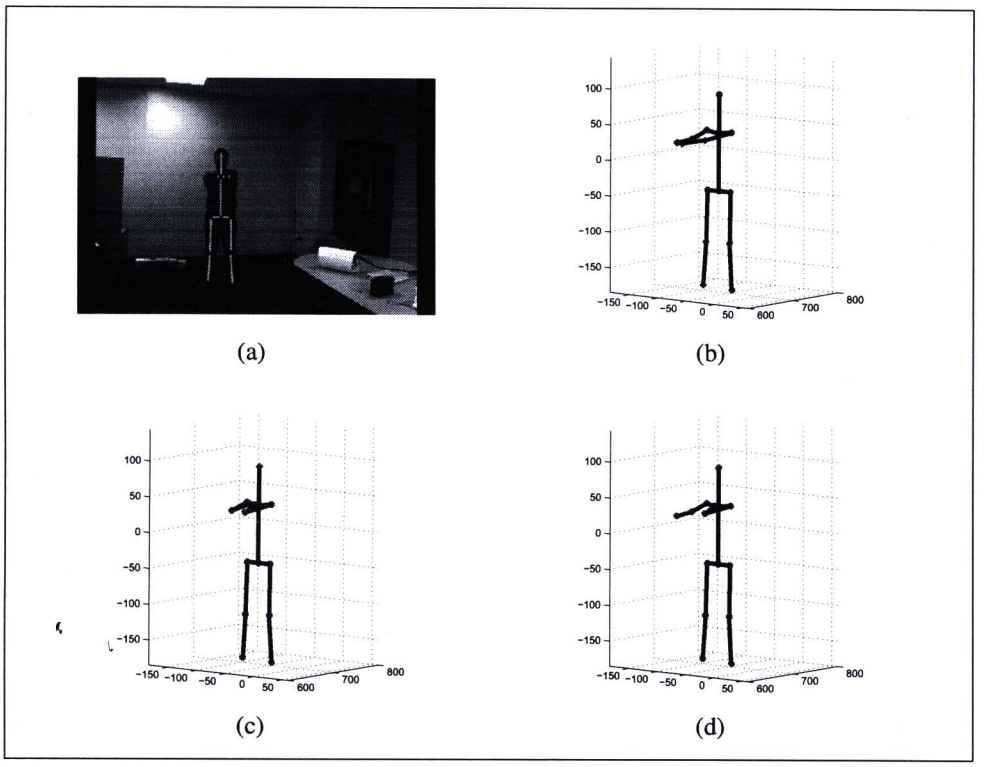

Figure 1.2 Problem of non-uniqueness of solutions in 3D human body pose tracking using a single camera (a) 2D joint point on an input image and (b) - (d) examples of possible 3D human poses constructed

allows a maximum of three degrees of freedom (DOFs) per joint. In this sense, it leads to a high dimension state space (generally over 33 dimensions). The main concept of this generative approach is to generate samples and then to measure similarity. The computational complexity of the approach is $O(N^m)$, where N is the number of samples for each body part and m the number of body parts. In general, many samples are generated leading to a huge increase of computational time. Moreover, the generative approach still requires a good initialization. Figure 1.3 shows an example of human representation by a 3D kinematic tree. θ_i is the orientation of the i^{th} segment to its corresponding joint.

Figure 1.3 High dimensional space of human representation in 3D human body tracking from the generative approach

Self-occlusion In monocular approach, multiple human body parts often occupy the same region in the input image due to its fixed viewing angle. Missing data is usually approximated by some prior knowledge of motion. Figure 1.4 shows example cases of self-occlusion that the right arm is occluded by the torso.

Observation Ambiguity Ambiguity arises from some symmetric body parts such as arms and legs that normally have similar appearances. Figure 1.4 shows examples of observation ambiguity of the right and left legs. It is difficult to identify correct part by a single camera.

Figure 1.4 Problems inherit the monocular approach: self-occlusion and observation ambiguity

Recovering from Lost Tracking In human body tracking applications, both multi-view and single view approaches generally face the same problem that is recovering from lost tracking. The problem is similar to (re)initialization state. Figure 1.5 shows an example of the lost tracking. Figure 1.5 (a) shows correct result of tracking. Figure 1.5 (b) - (d) show two lost tracking cases of the right leg. The lost tracking usually occurs after self-occlusion or observation ambiguity. Several approaches used prior knowledge in motion model to recover from lost tracking [70].

The focus of dissertation is on the development of a marker-less-based system for estimating and tracking 3D articulated human pose using a single camera. Like most of the monocular approach [21, 25, 28, 29], our work requires that all joint points are available a priori in the first frame. By the 3D human pose, we mean the relative 3D human skeletal model that is an up-to-scale version of the corresponding 3D real-word human pose. Unlike previous approaches, our proposed method does not require training or data set of 3D known pose exemplars, nor does it require the assumption that at least one predefined segment is parallel to the image plane or that most human parts are close to a plane parallel to the image plane in every frame. Our work, however, assumes a simpler assumption, for example, the actor stands vertically parallel to the image plane and not all of his/her joints lie on a plane parallel to the image plane in the first frame.

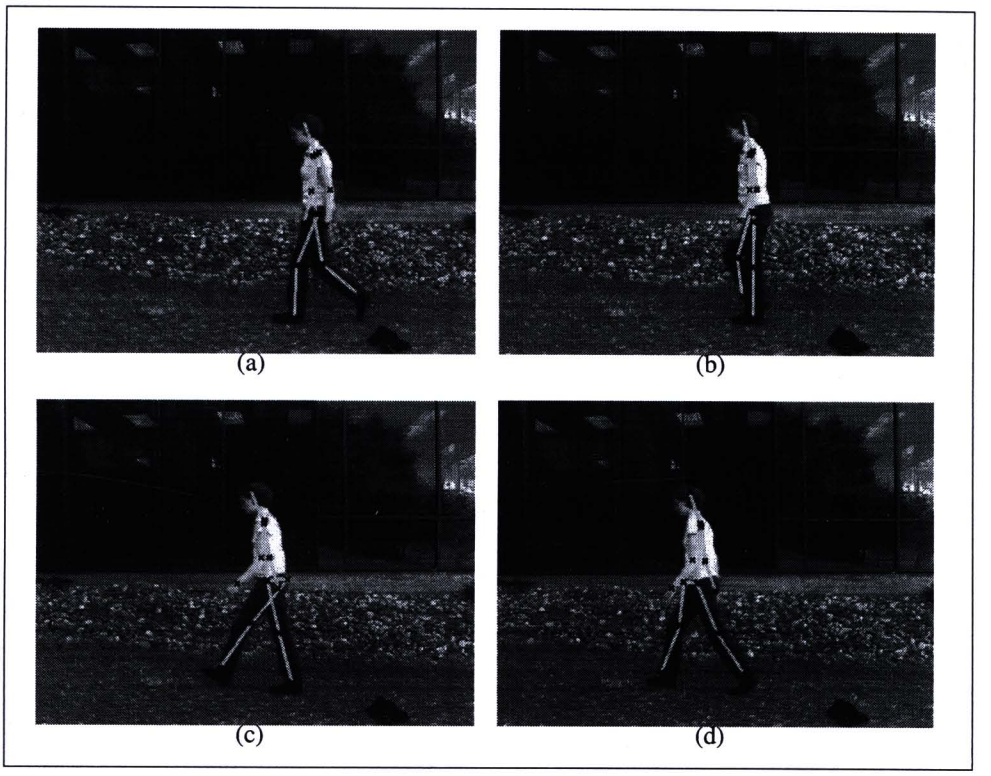

Figure 1.5 Example of the lost tracking (a) result of tracking before self-occlusion and observation ambiguity and (b)-(d) lost tracking of the right leg after self-occlusion and observation ambiguity

1.2 Objectives

The aim of this dissertation to develop a marker-less-based system for estimation and tracking 3D articulated human pose using an uncalibrated single camera.

1.3 Contributions

The basic idea of our approach is to reconstruct 3D relative human pose from 2D point correspondences. Our method consists of two main modules: 2D human body tracking and 3D articulated human reconstruction modules. The followings are our contributions:

1. A new approach for 2D human body tracking is proposed by integrating Quick Shift, a simple and efficient mode seeking method, into the belief propagation framework. It can reduce the computational complexity much more than the other methods due to the reduction of search space while preserving accuracy.

2. We apply feedback information from 3D human pose to alleviate several problems inherit in 2D human body tracking by a single camera, e.g. self-occlusion and observation ambiguity problems. Moreover, a motion model based on such feedback information and geometric constraint is introduced for good initializing state and (re)initializing state in case of lost tracking.

3. A novel approach to estimate and track 3D relative articulated body from 2D point correspondences is proposed. Unlike previous approaches, our proposed

method does not require camera parameters or a manual specification of the 3D pose at the first frame, nor does it require the assumption that most human parts are close to a plane parallel to the image plane, that at least one predefined human part is parallel to the image plane or that at lest one joint moves parallel to the image plane in every frame. It shows excellent results, especially in scenes with strong perspective effect.

4. We propose an efficient technique based on Multiple Hypothesis Tracking (MHT) to select the best configuration at the current frame using the past temporal information while still maintaining a number of most-likely previous pose trajectories. This method can alleviate the problem associated with the non-uniqueness of solutions. Moreover, it can recover lost tracking due to incorrect decision in early frames.

1.4 Dissertation Outline

Chapter 1 Introduction A general introduction of human body tracking are described. The chapter also introduces challenging problems of 3D human body tracking using a monocular image sequence. Moreover, contributions of this dissertation are presented in this chapter.

Chapter 2 Literature review In this chapter, we explain approaches in 3D human estimating and tracking, both top-down and bottom-up approaches. It also covers human model representation and image feature extraction. Additionally, this chapter explains approaches for 3D human model inference from either a single or multiple camera(s).

Chapter 3 Mathematical Background This chapter briefly review mathematical background used in our approach. It covers graphical model and belief propagation inference approaches. Moreover, some mode seeking techniques are explained.

Chapter 4 Overview of the proposed method This chapter introduces system overview of this dissertation. In addition, the chapter also introduces assumption and human model in our approach.

Chapter 5 2D human body estimation and tracking In this chapter, we present our proposed approach for 2D efficient human body tracking called Quick Shift Belief Propagation (QSBP) in a monocular image sequence. This chapter demonstrates the use of QSBP with a motion model based on feedback information from 3D human pose and and geometric constraint. In our experiments, we show results of 2D human body tracking under self-occlusion and observation ambiguity cases. Moreover, it presents improved performance obtained from reducing computational time by comparison with other works.

Chapter 6 Reconstruction and tracking of 3D-articulated human body from 2D point correspondences of a monocular image sequence This chapter introduces a new perspective approach for reconstruction of 3D articulated human body from an uncalibrated monocular image based on a perspective camera model. The chapter shows performance of this approach by evaluation on both synthesized and real-world image

sequences. In addition, it also introduces an approach to find the optimal solution due to the non-uniqueness of solutions and also shows robustness and comparison with other approaches.

Chapter 7 Discussions and Conclusions Finally, we conclude and discuss performance of our proposed method. It also covers the suggestions.