CHAPTER 5 2D HUMAN BODY ESTIMATING AND TRACKING

The 2D human body tracking is the first module in our approach as mentioned in Section 4.1. An image sequence taken from an uncalibrated camera served as its input and then all 2D human joint points are obtained for reconstruction of 3D relative human body pose in the second module. In this chapter, we present our Quick Shift Belief Propagation (QSBP) based approach which benefits from Quick Shift, a simple and efficient mode seeking method, in a part based belief propagation model. It can reduce computational time due to convergence with linear complexity in the number of body parts. Moreover, our approach performs with local samples and weights instead of evaluating all the possible states. From this concept, it needs a significantly smaller number of samples than other part-based approaches [81, 28, 100, 101, 102, 103, 80, 43, 104] do leading to a great saving of computational time.

Additionally, the unique aspect of Quick Shift model is its ability to efficiently discover modes of the underlying marginal probability distribution while preserving the accuracy. This gives our QSBP a significant advantage over approaches like Belief Propagation (BP) and Mean Shift Belief Propagation (MSBP) [29]. A motion model based on feedback information from the 3D human pose estimation and geometric constraint is introduced for good initializing state and reinitializing state in case of lost tracking. Moreover, we apply such feedback information to alleviate several problems inherit in 2D human body tracking using a single camera, e.g. self-occlusion and observation ambiguity problems. In experiments, we present qualitative and quantitative analysis of the proposed approach with encouraging results.

This chapter is organized as follows. Section 5.1 describes related work. MSBP [29] is explained in Section 5.2. Section 5.3 explains the proposed 2D articulated body human tracking. Experimental results and evaluations are presented in Section 5.4. Finally, conclusions are drawn in Section 5.5.

5.1 Related Work

Currently, the part-based approach is a very popular technique for articulate human body tracking to deal with the high dimensionality problem. It is more powerful than the generative approach (top-down approach) due to reducing computational cost from $O(N^m)$ to $O(mN^2T)$ [29, 108], where N is the number of samples for each body part, m is the number of body parts and T is the number of iterations needed for convergence. Instead of exponential complexity as the generative approach, the part-based approach converges with linear complexity in the number of body parts.

In the part-based approach, the human body is represented by a graphical model where each part is described by a node and the relationship between the adjacent parts is indicated by an edge of the graph. Each body part is considered separately and then is combined into the global solution later. The main idea of the part-based approach is based on message sending between adjacent nodes of the graph. The overview of part-based approach is shown in Figure 5.1. It is first performed by finding a set of can-

didate body parts by either sampling method [81, 28, 100, 101, 102, 103, 80, 43, 104] or generating sample locations [29]. Then, their beliefs are computed from observation functions and messages out to find an optimal solution.

Two popular techniques used to calculate the messages are the belief propagation [81, 28, 100, 101, 102, 103, 80, 43, 104, 29] and mean field Monte Carlo methods [111]. Their difference lies in the pattern of messages. In belief propagation, every node sends different messages to its neighbors. The message of each node is obtained from collection of messages received from all their neighboring nodes in the previous iteration. In mean field concept [111], every node send a single message to its neighbors based on the messages that is received from all of its neighboring nodes in the previous iteration. Shen at el. [28] compare and report that belief propagation offers better performance than that of the mean field Monte Carlo method.

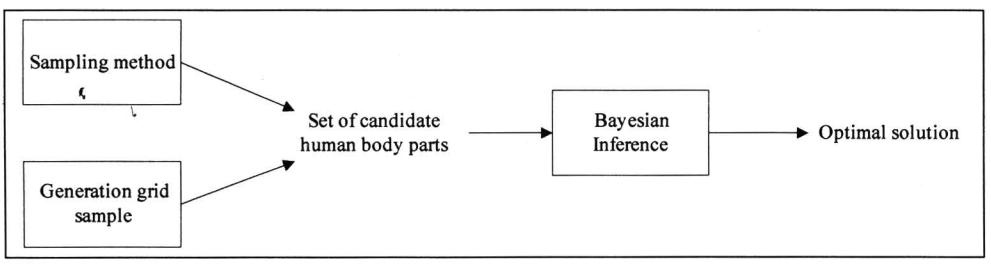

Figure 5.1 Overview of part-based approach

To tackle problems with high dimensional continuous state spaces, belief propagation technique with particle filter concept [62] are proposed. These include Nonparametric Belief Propagation algorithm (NBP) by Sudderth et al. [100] and Particle Message Passing algorithm (PAMPAS) by Isard [101]. Both algorithms extend the flexibility of particle filter concept [62] into the belief propagation framework. They approximate the message product of belief propagation by a mixture of Gaussians and then sampling samples are performed from them. However, there are two key differences between NBP [100] and PAMPAS [101]. First, the potentials of each node is formed by Gaussian and Gaussian outlier in PAMPAS [101], while no particular form for the potentials is assumed in NBP [100]. Moreover, an import sampling in NBP [100] is incorporated directly into Gibbs sampler [112], while the import sampling is based on weights of samples in PAMPAS [101]. By their effectiveness, both algorithms are widely used in high dimensional applications.

For example, Sudderth et al. model hand kinematics with a graphical model and use NBP [100] to track hand motion [102, 103]. Sigal et al. [43] present the general loose limbed body model to infer a 3D human pose with calibrated cameras. They address the problem of human pose estimation at a single time. To infer a 3D human pose, they employ the PAMPAS framework [101] and sample sampling with the Gibbs sampler [112] as in NBP approach [100]. To increase efficiency in tracking, they adopt a detector for initialization information in the first iteration of belief propagation. In [80], Sigal et al. extend their previous work [43] over time to perform 3D loose-limbed human tracking.

In [104], Sigal and Black introduce an extension of an approximate PAMPAS algorithm [101] to recover the real-valued 2D pose of the human body in the presence of occlusions. In their approach, occlusion-sensitive models are obtained by a learning process from 8 discrete views of at same person. They assume that the depth ordering of the human body parts is known a priori in each view. Although NBP has been a popular inference algorithm for graphical model with high dimensional problem, the NBP algorithm [100] is slow due to the sampling process. A more efficient version of NBP [100] is proposed using sequential density estimation and mode propagation [108]. They report that their method is 80 times faster than standard BP for tracking a 2D articulated body model; however, it still far from adequate for real-time tracking requirement.

To avoid complicating of the assumption for approximating intractable continuous belief propagation, several approaches utilize discretization, Discrete Belief Propagation (DBP). Hua et al. [81] adopt samples and their weights to approximate belief propagation. They model both the messages and marginal distributions with weighted samples. The message passing process is computed efficiently based on the samples drawn from an importance sampling built from the bottom-up approach, e.g. importance function for head pose, arm pose, leg pose and torso pose. Like Hua et al. [81], Shen et al. [28] adopt the import sampling concept for drawing samples. They propose a graphical model to extend across time frames in 2D articulated tracking. They recompute weights of each sample in drawing samples for the next frame.

Gao and Shi [24] propose multiple frame motion inference belief propagation to track upper human body parts. To guide the sample generation in order to achieve better efficiency in tracking, they detect human face and hands by color cues and approximate position of the shoulder based on the face position. To speed up the belief propagation, Ramanan and Forsyth [53, 113] reduce the size of input to the belief propagation framework by removing states that have low image likelihood value in 2D human body tracking applications. However, their method need additional pre-processing time for evaluation of image likelihood. Moreover, their pruning method is prone to error; once an important state is wrongly pruned in the pre-processing stage.

Instead of evaluating all the possible states, an efficient belief propagation method based on local samples is proposed by Park et al. [29]. They proposed Mean Shift Belief Propagation (MSBP). The results show that it is efficient and robust, and outperforms DBP and NBP [100] in terms of accuracy and stability. They report that MSBP [29] is 30-50 times faster in 2D state vector tracking (multi-target tracking), and 300 times faster in 3D state vector tracking than DBP. The concept of Park et al. [29] in work with local samples is an inspiration of 2D human body tracking in our proposed method. The details of the MSBP approach [29] are given in the next section.

5.2 Mean Shift Belief Propagation [29]

Mean Shift Belief Propagation (MSBP) is proposed by Park et al. [29]. The main concept is reducing time complexity by avoiding to construct probabilities of the entire

belief surface. They apply Mean Shift into the belief propagation framework to climb hills on the belief surface. Their approach works iteratively with local samples and their weights. In this concept, MSBP needs a significantly smaller number of samples than other approaches while it also yields accurate and stable solutions.

Firstly, A discrete grid of samples is generated around the predicted initial state of each node in the first iteration of belief propagation framework. Then, weight of each sample is computed by belief propagation concept as shown in equation (5.1). The new predicted states of each node are computed for hill-climbing belief surface to reach the optimal solution. To compute the new predicted state, a nonparametric mode-seeking Mean Shift technique is applied as shown in equation (5.3). A new set of samples is then generated around the recently predicted state and the weight is computed. The process repeats until convergence. In their approach, the optimal solution is used as the predicted initial pose for the next frame. The weights of node $\mathbf{x}_i$ at the n^{th} iteration can be computed by taking the product of incoming messages and local observations, as shown in equation (5.1).

$$b^n(\mathbf{x}_i) \leftarrow \alpha\phi(\mathbf{x}_i, \mathbf{z}_i) \prod_{j\in\Gamma(i)} m_{ji}^n(\mathbf{x}_i) \tag{5.1}$$

where $\mathbf{x}_i$ and $\mathbf{z}_i$ are the i^{th} hidden and corresponding observation nodes, respectively. $\phi(\mathbf{x}_i, \mathbf{z}_i)$ is the observation function of node i and α is a normalizing factor. From the belief propagation concept, the incoming messages also contain prior knowledge of the node obtained from the neighboring nodes as shown in equation (5.2). In the equation, $m_{ji}^n(\mathbf{x}_i)$ is a message sent from node j to node i at iteration n that can be calculated by

$$m_{ji}^n(\mathbf{x}_i) \leftarrow \sum_{\mathbf{x}_j} (\psi(\mathbf{x}_i, \mathbf{x}_j)\phi(\mathbf{x}_j, \mathbf{z}_j) \prod_{k\in\Gamma(j)\backslash i} m_{kj}^{n-1}(\mathbf{x}_j)) \tag{5.2}$$

where $\psi(\mathbf{x}_i, \mathbf{x}_j)$ is the potential function between nodes i and j, and $\Gamma(j)\backslash i$ represents all the neighboring nodes of node j except node i. From Mean Shift mode seeking on weighted samples, the updated position of node i at iteration $n+1$ is computed by

$$\mathbf{x}_i^{n+1} = \frac{\sum_{j=1}^{M} K(\mathbf{x}_i^n - \mathbf{s}_j) b^n(\mathbf{s}_j)\mathbf{s}_j}{\sum_{j=1}^{M} K(\mathbf{x}_i^n - \mathbf{s}_j) b^n(\mathbf{s}_j)} \tag{5.3}$$

where $b^n(\mathbf{s}_i)$ is the weight of the i^{th} sample, $\mathbf{x}_i^{n+1}$ is the estimated location of the initial predicted state of the i^{th} node after the n^{th} iteration and $\mathbf{s}_j$ is the j^{th} sample inside the Mean Shift kernel K.

5.3 Proposed Method

The main concept of our approach is applying Quick Shift mode seeking into belief propagation framework. Our approach can reduce time complexity while preserving the accuracy. It works with only local samples instead of evaluating all the possible states. By this concept, it needs a significantly smaller number of samples than other approaches in most of part-based approaches. Additionally, the unique aspect of our approach is its ability to efficiently discover modes of the underlying marginal probability distribution. It can converge faster than MSBP approach [29] because moving

toward modes of Quick Shift is based on the locally maximum probability value, while the Mean Shift is based on the weighted mean value.

In our QSBP approach, a motion-model-based on feedback information of 3D estimated human pose and a geometric constraint is introduced for good initializing state and reinitializing state in case of lost tracking. Samples are generated around the initial state and their probabilities are measured by the belief propagation. Only the best solution of each node is selected to be a part of the optimal pose solution. By applying Quick Shift with belief propagation for mode seeking in each body part, modes of the samples are obtained because all samples (initial values of the mode estimates) are not necessary moved to a single point. The modes are selected as initial samples for the next iteration of belief propagation. In this sense, it reduces risks of getting into spurious solutions which have a high probability. Moreover, we integrate spatial and temporal information into the observation computation for color and motion features to alleviate the observation ambiguity problem. Furthermore, we employ feedback information of 3D estimated human pose in generating a self-updated binary occlusion mask to alleviate several problems inherit in 2D human body tracking using a single camera, e.g. self-occlusion and observation ambiguity problems.

5.3.1 Human Representation

Human model is regarded as a graphical model with hidden nodes and pair-wise potentials as shown in Figure 5.2. The hidden nodes denote segments of human body. They are represented by $\mathbf{X} = \{\mathbf{x}_i | i \in [1, 14]\}$, where $\mathbf{x}_i$ is the i^{th} body part consisting of 4 states, i.e., position coordinate, orientation and length in the image plane. A corresponding observation set is denoted by $\mathbf{Z} = \{\mathbf{z}_i | i \in [1, 14]\}$, where $\mathbf{z}_i$ is the image observation node for the i^{th} body part. The relationship between $\mathbf{x}_i$ and $\mathbf{z}_i$ is represented by the observation function $\phi(\mathbf{x}_i, \mathbf{z}_i)$. In addition, every pair of adjacent body parts $\mathbf{x}_i$ and $\mathbf{x}_j$ (as defined by the structure), is connected and encoded by a potential function, $\psi(\mathbf{x}_i, \mathbf{x}_j)$. From Figure 5.2, the hidden nodes and and pair-wise potentials are shown by circles and squares, respectively. The arrows show the direction of message sending between adjacent nodes.

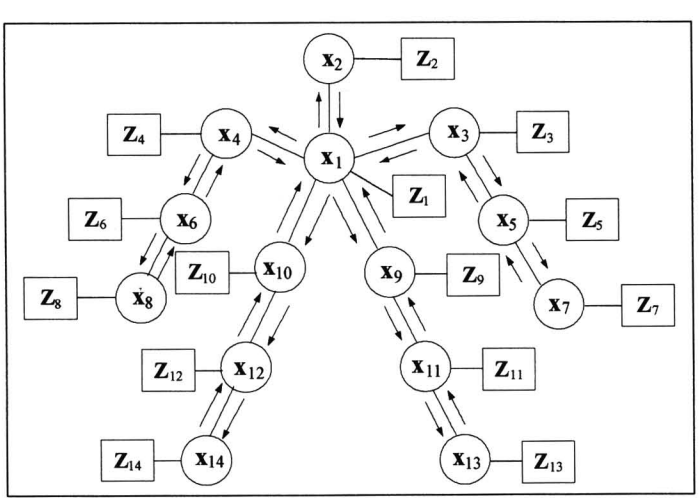

Figure 5.2 The graphical model in our approach

5.3.2 Motion Model

A motion model is utilized for better initializing state in our approach. The main concept is to employ feedback information of possible 3D human pose results obtained in the previous frames for the initialization state of the next frame . The 3D predicted configurations are evaluated as initial states and are projected onto the image plane for observation computation in our approach. We choose the first order motion model in the root joint for a starting point of predicting 3D human pose in the next frame, $t+1$.

$$\begin{bmatrix} x_1^{t+1} \\ y_1^{t+1} \\ z_1^{t+1} \end{bmatrix} = \begin{bmatrix} x_1^t \\ y_1^t \\ z_1^t \end{bmatrix} + \begin{bmatrix} v_{1,x}^t & 0 & 0 \\ 0 & v_{1,y}^t & 0 \\ 0 & 0 & v_{1,z}^t \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \nu \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

where (x_1^t, y_1^t, z_1^t) is the 3D virtual coordinate of the root joint at time t and ν is multi-variance of the white Gaussian noise. $v_{1,x}^t$, $v_{1,y}^t$ and $v_{1,z}^t$ are velocities of the root joint at time t along axes x, y and z, respectively.

For the other joints, they are predicted using their segment lengths and possible trajectory on the surface of a sphere. Suppose (x_p^t, y_p^t, z_p^t) is a known coordinate of the previous joint, the next joint (x_q^t, y_q^t, z_q^t) can be defined by

$$\begin{aligned} x_q^{t+1} &= x_p^{t+1} + l_{pq} \sin\omega_q^{t+1} \cos\psi_q^{t+1} \\ y_q^{t+1} &= y_p^{t+1} + l_{pq} \sin\omega_q^{t+1} \sin\psi_q^{t+1} \\ z_q^{t+1} &= z_p^{t+1} + l_{pq} \cos\omega_q^{t+1}, \end{aligned}$$

where l_{pq} is length of segment pq. ω_q^{t+1} and ψ_q^{t+1} can be computed by

$$\begin{bmatrix} \omega_q^{t+1} \\ \psi_q^{t+1} \end{bmatrix} = \begin{bmatrix} \omega_q^t \\ \psi_q^t \end{bmatrix} + \begin{bmatrix} \omega_q^t - \omega_q^{t-1} & 0 \\ 0 & \psi_q^t - \psi_q^{t-1} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \xi \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

where ξ is the variance of the white Gaussian noise. From the first two frames of the image sequence, ω_q^t and ψ_q^t are computed by

$$\begin{aligned} \omega_q^t &= \arctan\left(\frac{y_q^t - y_p^t}{x_q^t - x_p^t}\right) \\ \psi_q^t &= \arccos\left(\frac{z_q^t - z_p^t}{\sqrt{(x_q^t - x_p^t)^2 + (y_q^t - y_p^t)^2 + (z_q^t - z_p^t)^2}}\right). \end{aligned}$$

We have found from our empirical studies that a good tracking performance can be obtained if the predicted joint gives an initial state within 1.5 times of the grid size. For experiments, we initial state in the first iteration of belief propagation with ground truth and corrupt by Gaussian noise. 4D Gaussian distributions with zeros mean and variance 0.5, 1, 1.5 and 2 times of grid size are generated and used to be initial state in our QSBP approach. The average distance errors from the corresponding ground truth are 2.1, 2.17, 2.6 and 8.7 pixels/joint, respectively.

To increase performance of tracking, the variance is extended in self-occlusion case which is indicated by the estimated 3D human pose from the previous frame. In this sense, it can recover from 2D lost tracking that is similar to (re)initializing the state.

Figure 5.3 (a) shows 2D tracking result on the input frame at time t and 3D human model using our proposed method is displayed in Figure 5.3 (b). The 3D predicted configurations from motion model and their projected joints on the input image at time $t+1$ are shown in Figure 5.3 (c) and (d), respectively.

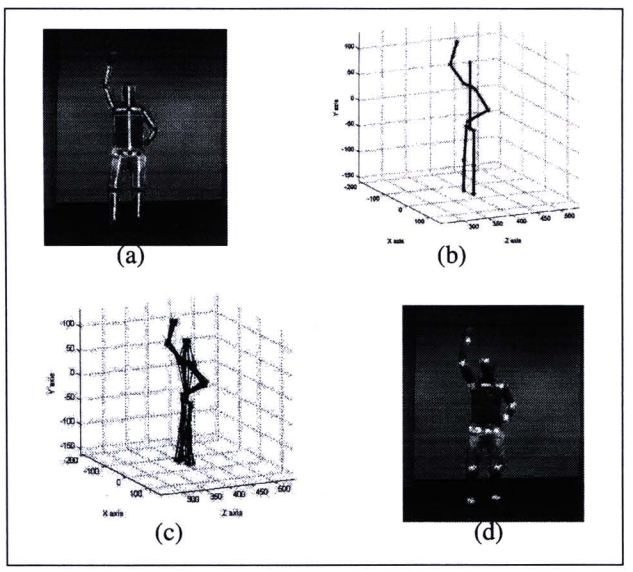

Figure 5.3 Our motion model process (a) input image at time *t*, (b) 3D articulated human body from our approach , (c) 3D predicted articulated human from motion model, (d) 2D projected points of 3D predicted articulated human on the input image at time *t+1*

5.3.3 Quick Shift Belief Propagation

The belief propagation algorithm is an iterative method to infer the hidden node of graphical model until it converges to the optimal solution. The marginal probability (belief) of hidden node $\mathbf{x}_i$ at iteration n, $p^n(\mathbf{x}_i|\mathbf{Z})$ can be computed by taking the product of incoming messages and local observations, as shown in equation (5.4). The incoming messages also contain prior knowledge of the node obtained from the neighboring nodes as shown in equation (5.5).

$$p^n(\mathbf{x}_i|\mathbf{Z}) \leftarrow \alpha\phi(\mathbf{x}_i, \mathbf{z}_i) \prod_{j\in\Gamma(i)} m_{ji}^n(\mathbf{x}_i). \tag{5.4}$$

where $\mathbf{x}_i$ and $\mathbf{z}_i$ are the i^{th} hidden and corresponding observation nodes, respectively. $\phi(\mathbf{x}_i, \mathbf{z}_i)$ is the observation function of node i. α is a normalizing factor. For graphical models with continuous hidden state, $m_{ji}^n(\mathbf{x}_i)$ is a message sent from node j to i at iteration n and can be calculated by

$$m_{ji}^n(\mathbf{x}_i) \leftarrow \alpha \int_{\mathbf{x}_j} (\psi(\mathbf{x}_i, \mathbf{x}_j)\phi(\mathbf{x}_j, \mathbf{z}_j) \prod_{k\in\Gamma(j)\backslash i} m_{kj}^{n-1}(\mathbf{x}_j))d\mathbf{x}_j. \tag{5.5}$$

where $\psi(\mathbf{x}_i, \mathbf{x}_j)$ is the potential function between nodes i and node j, $\phi(\mathbf{x}_j, \mathbf{z}_j)$ is the observation function of node j and $\Gamma(j)\backslash i$ represents all neighboring nodes of node j except node i.

In our approach, we define parts of human body by nodes of the graph that the optimal hidden node is computed by maximizing the marginal probability of each node given the current observation in the graphical model as shown in Figure 5.2. Instead of considering all possible states of our nodes, we work on a grid around initial samples (modes found in the previous iteration). A new discrete grid is generated around the modes. The grid size is 5x5x3x3 of 4 states (x-y positions, length and angle of the body part). The marginal probability of node $\mathbf{x}_i$, $p(\mathbf{x}_i|\mathbf{Z})$, can be computed by taking the product of incoming messages and local observations as shown in equation (5.4). The incoming messages also contain prior knowledge of the node obtained from its neighboring nodes. The message sent from node j to node i at iteration n, $m_{ji}^n(\mathbf{x}_i)$, can be calculated by

$$m_{ji}^n(\mathbf{x}_i) \leftarrow \alpha \sum_{\mathbf{x}_j} (\psi(\mathbf{x}_i, \mathbf{x}_j)\phi(\mathbf{x}_j, \mathbf{z}_j) \prod_{k \in \Gamma(j) \setminus i} m_{kj}^{n-1}(\mathbf{x}_j)). \tag{5.6}$$

The message at the first iteration, $m_{ji}^0(\mathbf{x}_i)$, is defined to be 1. From the iterative concept of belief propagation, the best solution is obtained in the final iteration by

$$\mathbf{x} = \underset{\mathbf{s}_i \in \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N\}}{\arg\max} \; p^n(\mathbf{s}_i|\mathbf{Z}), \tag{5.7}$$

where N is the number of samples inside of a grid window.

5.3.4 Mode Seeking in Belief Propagation

In mode seeking, we use marginal probability of belief propagation in movement of mode estimates. Only samples around the modes are computed, not the entire surface of belief propagation. This makes the convergence to an optimal solution very fast and computation time is reduced. The updated position of sample $\mathbf{s}_i$ at iteration $k+1$ of the mode seeking is computed by

$$\mathbf{y}_i^{k+1} = \underset{\mathbf{s}_j \in \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N\} : P(\mathbf{s}_j|\mathbf{Z}) > P(\mathbf{y}_i^k|\mathbf{Z}), D(y_i^k, \mathbf{s}_j) < \kappa}{\arg\min} D(\mathbf{y}_i^k, \mathbf{s}_j), \tag{5.8}$$

where $D(\mathbf{y}_i^k, \mathbf{s}_j)$ is the distance between the current position of $\mathbf{y}_i^k$ and the i^{th} sample which is less than a distance threshold κ and $P(\mathbf{s}_i|\mathbf{Z})$ is the marginal probability value of the i^{th} sample. N is the number of samples in the grid. From Quick Shift mode seeking concept, the position of $\mathbf{y}_i$ is updated until no further change in labelling occurs and then modes of such samples are obtained as a unique set

$$mode = \{\mathbf{y}_i^t\}, \tag{5.9}$$

where $\mathbf{y}_i^t$ is the final position of $\mathbf{y}_i$.

We find that searching for modes of marginal probability in belief propagation by our approach which is faster than by MSBP [29]. Because moving toward modes of Quick Shift concept is based on the locally maximum probability value, while the Mean Shift concept is based on the weighted mean value. In this sense, the number of iterations in moving to modes of Quick Shift concept is less than that of Mean Shift cocnept which

makes our QSBP approach converge to the optimal solution much faster than MSBP [29]. Figures 5.4 (a) and (b) show convergence to the optimal solution by QSBP and MSBP [29], respectively. From these figures, each initial sample is plotted by a solid square marker. The only grid members (plotted by markers inside of a grid window) around the initial sample are considered in moving toward a mode of the sample. The new position of the mode estimate in each iteration until they reach a mode is shown by a circle marker.

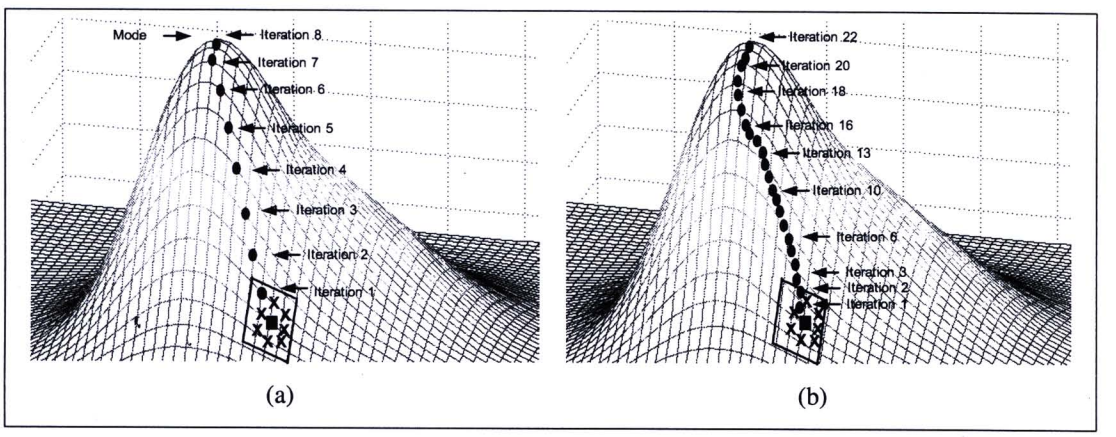

Figure 5.4 A comparison of moving toward the optimal solution by (a) our QSBP approach, and (b) MSBP approach [29]

Main steps of proposed approach

1. Generate initial joint predictions from a motion model as explained in Section 5.3.2.
2. Iterate steps 3-6 until convergence.
3. Generate a local grid of samples around each initial sample.
4. Compute message by equation (5.6).

$$m_{ji}^{n}(\mathbf{x}_i) \leftarrow \sum_{\mathbf{x}_j}(\psi(\mathbf{x}_i,\mathbf{x}_j)\phi(\mathbf{x}_j,\mathbf{z}_j) \prod_{k\in\Gamma(j)\setminus i} m_{kj}^{n-1}(\mathbf{x}_j)).$$

5. Compute marginal probability using in equation (5.4).

$$p^{n}(\mathbf{x}_i|\mathbf{Z}) \leftarrow \alpha\phi(\mathbf{x}_i,\mathbf{z}_i) \prod_{j\in\Gamma(i)} m_{ji}^{n}(\mathbf{x}_i).$$

6. Seek mode using Quick Shift, set each mode as the initial sample for the next iteration.
7. Select the best solution as

$$\mathbf{x} = \underset{\mathbf{x}_i\in\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\}}{\arg\max} \; p^{n}(\mathbf{x}_i|\mathbf{Z}),$$

5.3.5 Observation Function

Generally, the observation function $\phi(\mathbf{x}_i,\mathbf{z}_i)$ is used to measure the segment likelihood of $\mathbf{z}_i$ and $\mathbf{x}_i$. We integrate spatial and temporal information into the observation computation and apply a binary occlusion mask to an input image. They can alleviate detection of spurious color feature and observation ambiguity problem. In appearance information, we employ color feature. Each body part $\mathbf{x}_i$ is modeled by a planar patch and projected onto the input image, and then its likelihood (or similarity) is computed. We use hue (H) and saturation (S) components of the HSI color space [51] to construct the color feature. Lightness (I) is not used in the observation function so that it is robust against global illumination changes. The color feature of each human body is formed by a Gaussian mixture model $G(\mathbf{x}_i)$ based on HS components.

$$G(\mathbf{x}_i) = \sum_{j=1}^{M} w^j \eta(\mathbf{x}_i;\mu^j,\wedge^j) \tag{5.10}$$

where $\eta(\mathbf{x}_i;\mu^j,\wedge^j)$ denotes the j^{th} Gaussian component with mean μ^j and covariance $\wedge^j$. w^j is the j^{th} mixing weight satisfying $\sum_{j=1}^{M} w^j = 1$. From the color models of the i^{th} node, $G(\mathbf{x}_i)$, and the template of node i, $C(i)$, the similarity measure is defined by the distribution intersection between the color model of the sample and the template as

$$w_c = G(\mathbf{x}_i) \cap C(i) \tag{5.11}$$

where w_c is the color score. (A discrete approximation of the intersection is implemented in our work.)

For the temporal feature, we also project the 3D human pose from the motion model into the input image. The 2D projected predicted joint positions are modeled by a Gaussian model. The distance score to the predicted point of each sample is measured by Mahalanobis distance.

$$w_m = \frac{1}{\sqrt{2\pi\omega^2}} e^{-\frac{d^2}{2\omega^2}} \tag{5.12}$$

where w_m is the distance score, d is the Mahalanobis distance between sample and predicted positions, w_m the motion score and ω the standard deviation.

The observation function $\phi(\mathbf{x}_i, \mathbf{z}_i)$ of node $\mathbf{x}_i$ is computed by integrating of two features (color and motion features) as shown in equation (5.13).

$$\phi(\mathbf{x}_i, \mathbf{z}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(1-w_{c,m})^2}{2\sigma^2}} \tag{5.13}$$

where σ is the standard deviation. For a low value of σ, a more weight is given to the appearance similarity. $w_{c,m}$ is the observation value obtained by combining the color and the motion distance measures. It can be computed from

$$w_{c,m} = \alpha w_c + (1 - \alpha) w_m \tag{5.14}$$

where w_c and w_m are the color appearance and the motion scores, respectively. (The scores are normalized to be between 0 and 1.) α is the weight of blending between the color and the motion information.

5.3.6 Potential Function

The potential function $\psi(\mathbf{x}_i, \mathbf{x}_j)$ is used to represent the relationship between body parts i and j. We model the potential function by a Gaussian to represent the distribution of the Euclidean distance between the two adjacent body parts as shown in Figure 5.5.

$$\psi(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{\sqrt{2\pi\gamma^2}} e^{-\frac{D(\mathbf{x}_i, \mathbf{x}_j)^2}{2\gamma^2}} \tag{5.15}$$

where $\mathbf{x}_i$ is the i^{th} body part, $\psi(\mathbf{x}_i, \mathbf{x}_j)$ is a potential function of body parts i and j, $D(\mathbf{x}_i, \mathbf{x}_j)$ the Euclidean distance between the connected points of body parts i and j , and γ the standard deviation. In the same way as σ, γ specifies insensitivity to displacement of adjacent body parts.

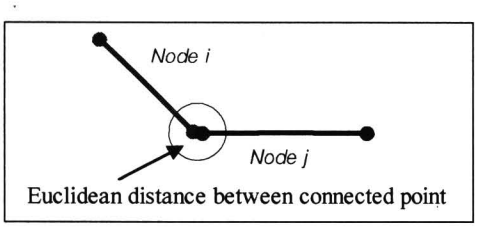

Figure 5.5 Euclidean distance between the connected points of body parts i and j

5.3.7 A Binary Occlusion Mask Determination

In each iteration of belief propagation in our approach, a binary occlusion mask of each part is first determined according to a human body part order. It is generated on 2D result obtained from the previous iteration. For the first iteration, the binary occlusion masks are based on 2D result from the previous frame. To obtain the ordering of human body part, we assume that distance from the center of the body part in 3D human pose (resulted from the previous frame) to the origin signifies the order of human body parts. Different binary occlusion masks are shown in Figure 5.6. A frame of a walking sequence is displayed in Figure 5.6 (a). In this figure, the obtained order of binary masks is the lower left arm, upper left arm, lower left leg, upper left leg, head, torso, lower right leg, upper right leg, lower right arm, upper right arm, accordingly. The binary occlusion mask for upper left arm and lower left leg tracking are shown in Figures 5.6 (b) and (c), respectively.

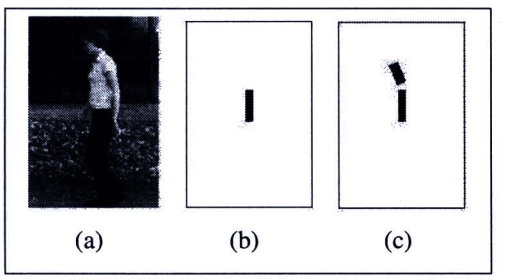

Figure 5.6 Different binary occlusion masks (a) an input image, (b) binary occlusion mask for the upper left arm tracking, and (c) binary occlusion mask of the lower left leg tracking

5.4 Experiments

In our experiments, We tested our approach with motion model in image sequences for both simple and occlusion cases. We compared our approach with DBP approach and MSBP [29] in several image sequences. Moreover, we compared our approach with MSBP [29] in 1D.

5.4.1 Performance of 2D Tracking with Motion Model

To evaluate the performance of our proposed method, we tested it on image sequences of UCF dataset [49] and walking action [114]. In our 2D human body tracking experiments, we generated a 5x5x3x3 grid and ran the algorithm until convergence (either joint position movement of less than 2 pixels or 50 iterations reached.) The distance threshold parameter in mode seeking of Quick Shift was chosen as half of the grid size. In the observation function, the standard derivation of observation function and potential function were chosen as 0.4 and 3, respectively. We assumed that the human actor stood vertically parallel to the image plane and not all of his/her joints lie on a plane parallel to the image plane, and that the joint locations were manually initialized in the first frame [21, 25, 28, 29]. Some results of 2D human body and 3D human pose reconstruction from two aerobic-styles and a walking scene using our approach are displayed in Figure 5.7 (a), (b) and (c), respectively. It can be seen that the model fits well to the images for all sequences.

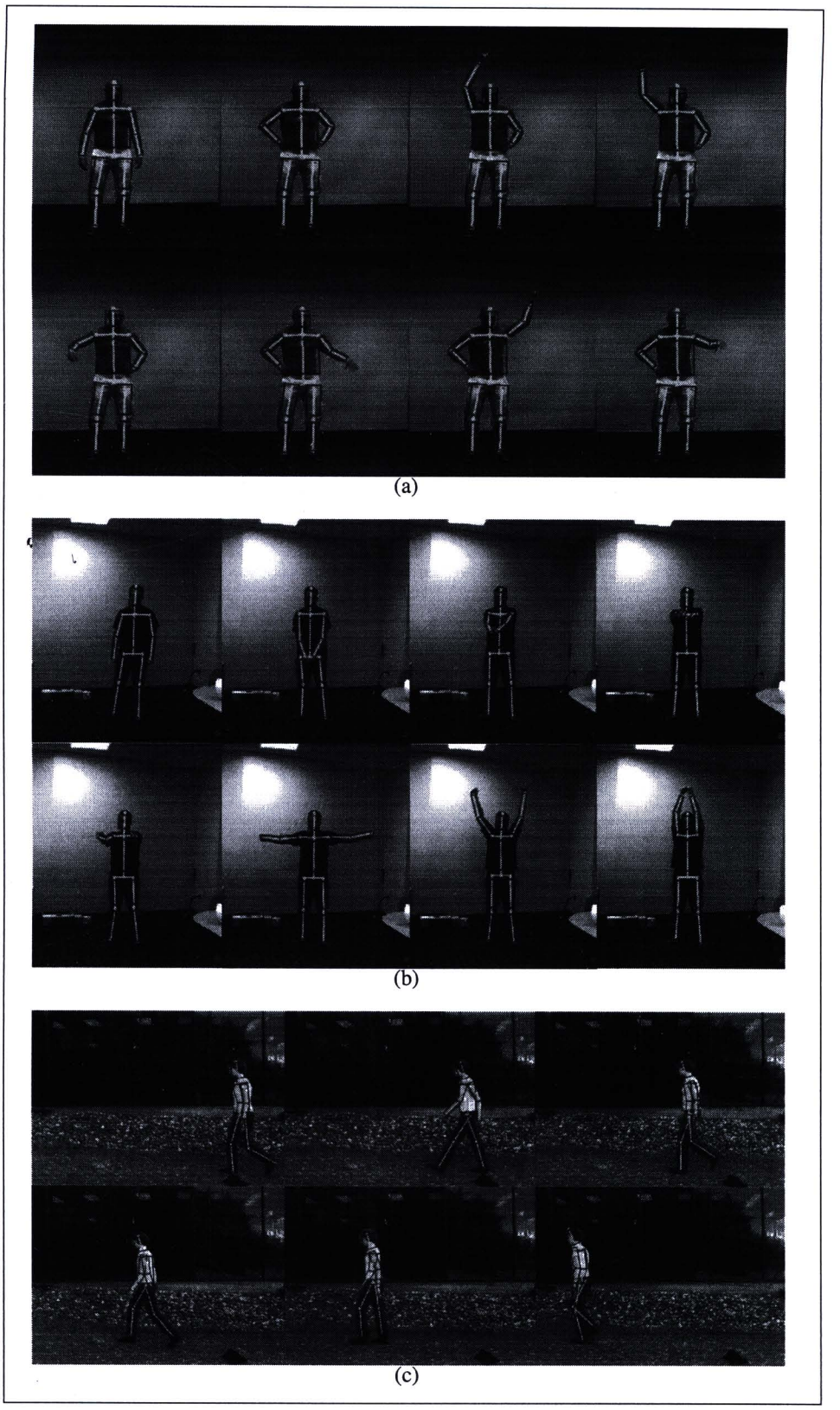

Figure 5.7 Sample of results in 2D human body tracking (a)-(b) the first and second image sequences of aerobics style activities [49], and (c) walking-scene image sequence [114]

5.4.2 Performance Comparison with MSBP [29] in 1D

We have experimented and showed the mode seeking by MSBP [29] and our QSBP approaches using various grid sizes in Figures 5.8 and 5.9, respectively (the grid size is twice the kernel sizes in Mean Shift). The bimodal curve in each image is the underlying marginal belief probability shown in 1D. The mode seeking by MSBP using grid sizes 4, 8, 12 and 16 is shown in Figures 5.8 (a) - (d), respectively. The initial center of grids and mode estimate (of each iteration) are shown by star and dot markers, respectively. The final mode estimate is shown by a square marker. It can be seen that MSBP [29] converges to a mode in every image; however, the kernel size effects the number of iterations required. The number of iterations used by MSBP [29] with kernel size 2, 4, 6 and 8 are 141, 57, 32 and 22, respectively. The Mean Shift with larger kernel size takes fewer number of iterations than that with smaller kernel size.

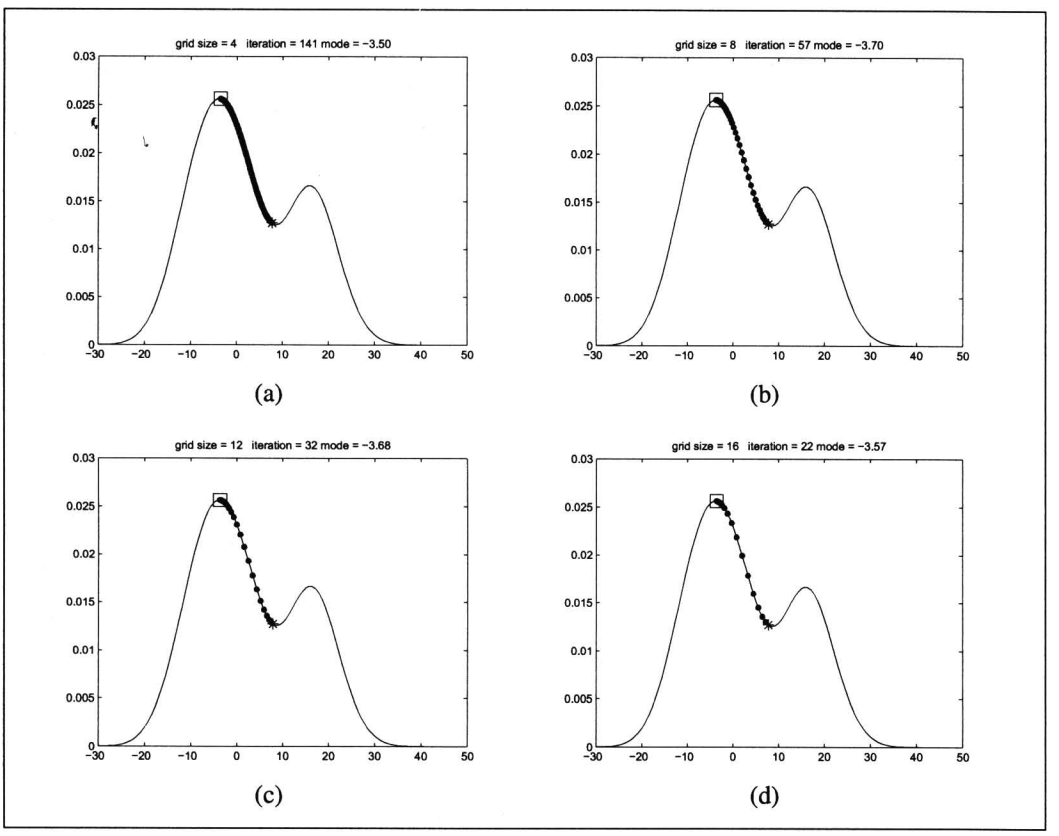

Figure 5.8 Mode seeking by MSBP (a) grid size = 4, (b) grid size = 8, (c) grid size = 12 and (d) grid size = 16

The mode seeking by QSBP using κ (equivalent to kernel size) of 4, 8, 12 and 16 is shown in Figures 5.9 (a) - (d), respectively. The initial center of grids and mode estimate (of each iteration) are shown by star and dot markers, respectively. The final mode estimate is shown by a square marker. It can be seen that QSBP gets into both modes and within 6, 3, 2 and 2 iterations, respectively. Our QSBP approach is an aggressive approach so that it can move to modes faster than MSBP approach [29].

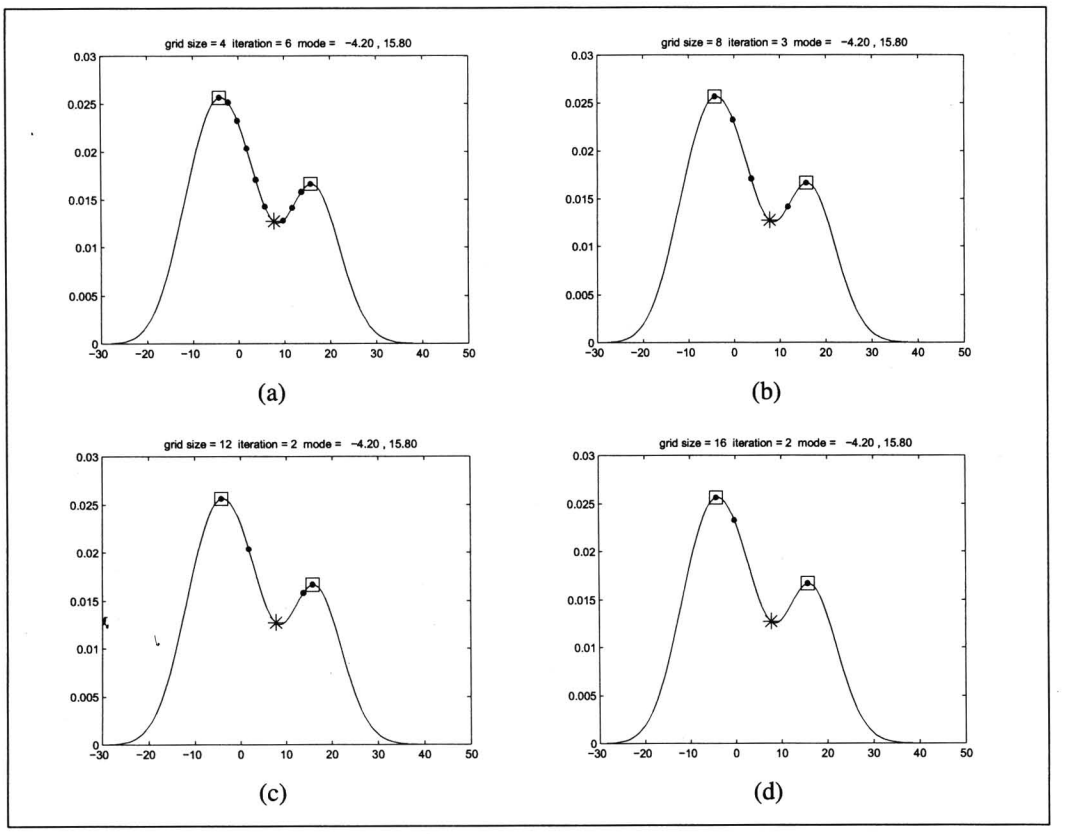

Figure 5.9 Mode seeking by QSBP (a) grid size = 4, (b) grid size = 8, (c) grid size = 12 and (d) grid size = 16

5.4.3 Performance Comparison with Other Approaches

To evaluate the performance of our proposed method, we tested it on several videos and compared its result with those from the BP and MSBP [29] methods. For the same environment, we applied a model video [49] to the sample prediction in the first iteration of belief propagation. The model video was first proposed by Gritai et al. [49]. They assumed that all joint points of human body are available for the entire model video but they are available for only the first frame of the test video. The main concept of the model video is to estimate the joint locations in the test video automatically using two geometric constraints. The affine constraint is used to estimate initial positions based on invariance ratio between model video and test video. Moreover, the epipolar constraint is used to reduce estimation error of different actors and view points between the model and test videos.

One advantage of this work is the avoidance of error propagation from frame to frame in the estimation process, because each joint estimate is computed based on correspondence between the first frame of the model and the test videos. Another advantage is robustness against variations in anthropometry, execution rate, viewpoint and execution style. Moreover, it is not computationally expensive and does not require extensive training. Figure 5.10 (a) shows the input frame. The candidate joints from the model video [49] are generated and overlaid on the input image as displayed in Figure 5.10 (b). The best match with the silhouette is selected to be the initial samples of belief propagation that can be seen in Figure 5.10 (c) for comparison our approach with BP

and NBP .

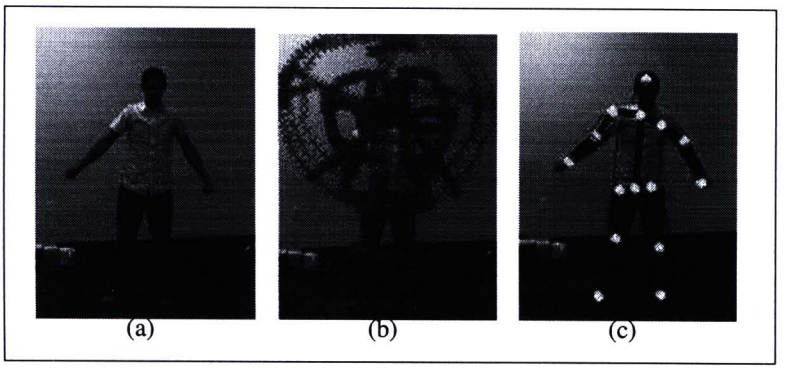

Figure 5.10 Prediction by model video (a) an input image, (b) predicted points from model video, and (c) initial configuration

In comparison, region overlapping [50, 48] and RGB color are the features used for similarity measurement in an observation function. The region overlapping feature of node $\mathbf{x}_i$ is computed by

$$w_i = \frac{1}{2}\left(\frac{n_{i,o}}{n_{i,p}} + \frac{n_{i,o}}{n_m}\right), \quad n_m = \arg\max_i \; n_{i,o} \tag{5.16}$$

where $n_{i,o}$ is the number of pixels in overlapping region between projected region of i^{th} node and the silhouette. $n_{i,p}$ is the number of pixels in the i^{th} projected region. This measure is very simple and efficient; however, its reliability reduces greatly in case of occlusion. This is because overlapped parts form a larger foreground region which provides high values of this feature in several false locations.

In case of occlusion, we therefore switch to using a more detailed color feature. For the color feature, each human body part is formed by a color histogram based on RGB components. From the color models of the node $H(\mathbf{x}_i)$ and the template, $C(i)$, the similarity measure is defined by the histogram intersection between the color shape matching model of the i^{th} node and the template as

$$w_i = \sum_{j=1}^{n} H_j(\mathbf{x}_i) \cap C_j(i), \tag{5.17}$$

where $H_j(\mathbf{x}_i)$ is the normalized number of pixels in the j^{th} bin of the i^{th} node and $H_j(i)$ is the normalized number of pixels in the j^{th} bin of template of node i in RGB color histogram and n is the number of bins in each histogram. The observation function $\phi_i(\mathbf{x}_i, \mathbf{z}_i)$ of sample $\mathbf{x}_i$ is computed by

$$\phi_i(\mathbf{x}_i, \mathbf{z}_i) = \frac{1}{\sqrt{2\pi\nu^2}} e^{-\frac{(1-w_i)^2}{2\nu^2}}, \tag{5.18}$$

where w_i is the similarity of node $\mathbf{x}_i$ obtained by equation (5.16) or Eq. (5.17) and ν is its standard deviation. For a low value of ν, a more weight is given to the appearance similarity.

To handle occlusion, we use two measures for detecting start and end of self-occlusion as in [49]. The first measurement is α_j^t, which represents the area of the foreground silhouette, corresponding to the j^{th} segment in the t^{th} frame. The second measure is β_j^t, which represents the proportion of the detected segment j that is occluded by the other segments of the cardboard model. The condition for occlusion is based on the normalized change over time τ,

$$\frac{\sum_{i=t-\tau}^{t-1} \alpha_j^{i+1} - \alpha_j^i}{\tau \alpha_j^{t-\tau}} < T \quad , \quad \frac{\sum_{i=t-\tau}^{t-1} \beta_j^{i+1} - \beta_j^i}{\tau \beta_j^{t-\tau}} > T. \tag{5.19}$$

where T is the percentage threshold and τ is the number of previous frames that are used to consider occlusion. Positive T value signifies occlusion entering, while the negative T value indicates occlusion termination. To select the size of temporal windows τ, it relates with the number of frames from starting occlusion to complete occlusion . For T value, it specifies sensitivity to the size of temporal windows, τ.

We tested our method on six sequences of UCF dataset, containing 400, 162, 400, 560, 500 and 500 frames. These videos include aerobic style activities that were also used in [49]. Moreover, we experimented on two sequences of CMU dataset (also used in [29]), containing 199 and 190 frames including a walking action in both front and side views. These videos include both simple and self-occlusion cases. To present qualitative and quantitative results, we compared our approach with [29]. Note that the joint locations were manually initialized in the first frame, and then an RGB template of each human part is automatically generated from the initialized joint locations. The sample prediction was then performed automatically for the remaining frames like the method presented in [49].

In our experiments, we generated samples with a size of 8,000 samples for BP, and a 5x5x3x3 grid for both MSBP [29] and QSBP, respectively. Those methods were run until convergence (either joint position movement of less than 2 pixels or 50 iterations reached.) The threshold parameter in mode seeking of Quick Shift and the kernel size of MSBP [29] were chosen as half of the grid size. Like [49] for occlusion handling, we used 70 percent in our experiments and the number of pervious frames was chosen as 15 to consider occlusion.

Samples of tracking results are displayed in Figure 5.11. The figure shows samples of human model fitting on four sequences. The fitting results of QSBP, BP and MSBP approaches [29] are shown in Figures 5.11 (a), (b) and (c), respectively. Since similar results among all methods were obtained, only two results were included for BP and MSBP [29] in Figures 5.11 (b) and (c). It can be seen that the model fits well to the images for each approach. The results show that the accuracy of QSBP is comparable to those of BP and MSBP [29] as illustrated in Figure 5.12. It is compared with ground truth which is manually obtained. The average distance errors from the corresponding ground truth are shown in Figure 5.12 (a). It can be seen that all methods provide accuracy; however, our approach is far more efficiency than the others as shown in Figure 5.12 (b). In particular, for our case of 4-state (2D position, body part length and angle) tracking, our proposed technique is respectively 36 and 99 times faster than those of

MSBP [29] and BP. On average, the numbers of iterations to get the best solution are 9, 30 and 42 for QSBP, MSBP [29] and BP, respectively.

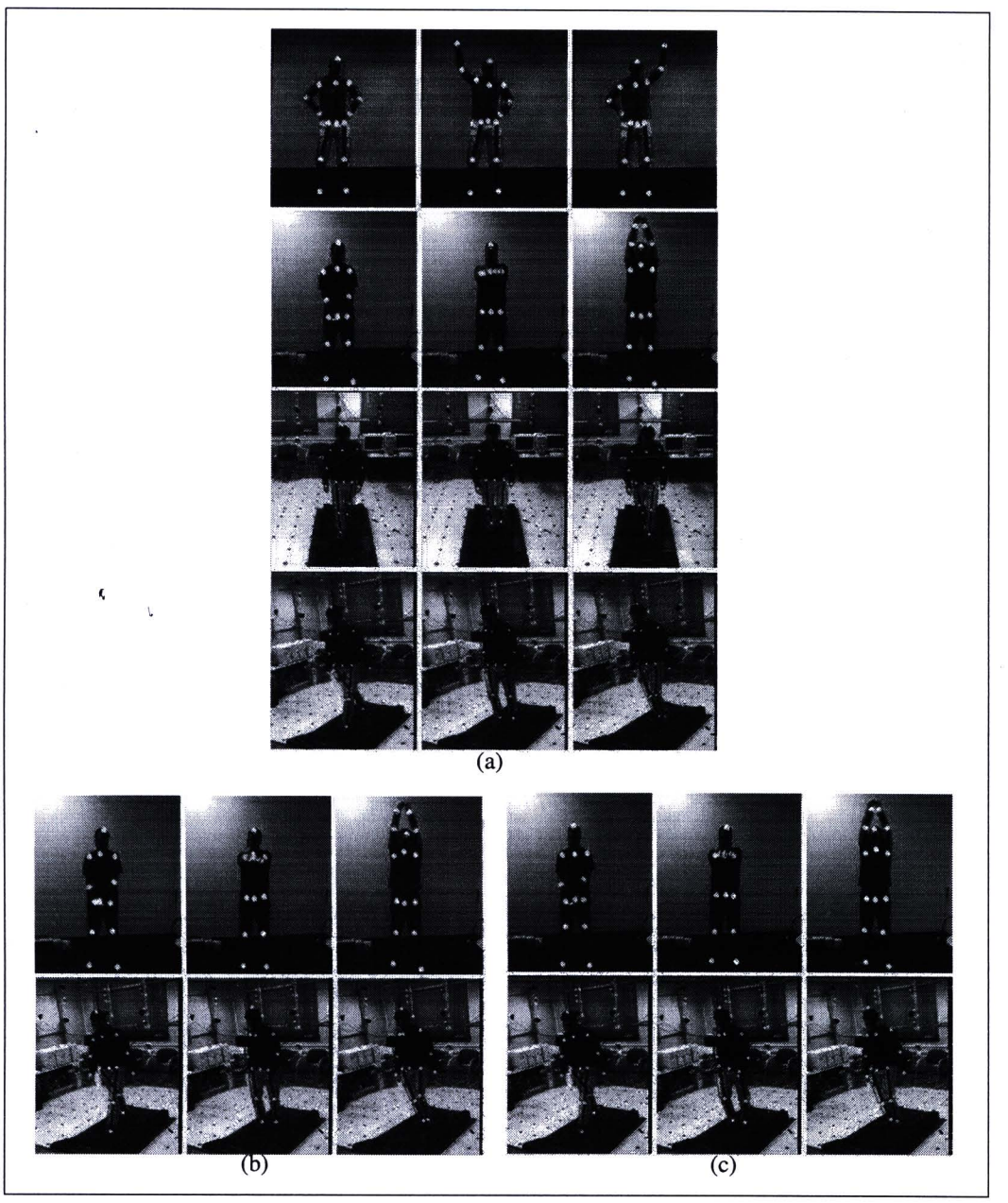

Figure 5.11 Some results of human body tracking by using model video in sample prediction. Similar results from (a) Quick Shift belief propagation, (b) Belief propagation and (c) Mean Shift propagation

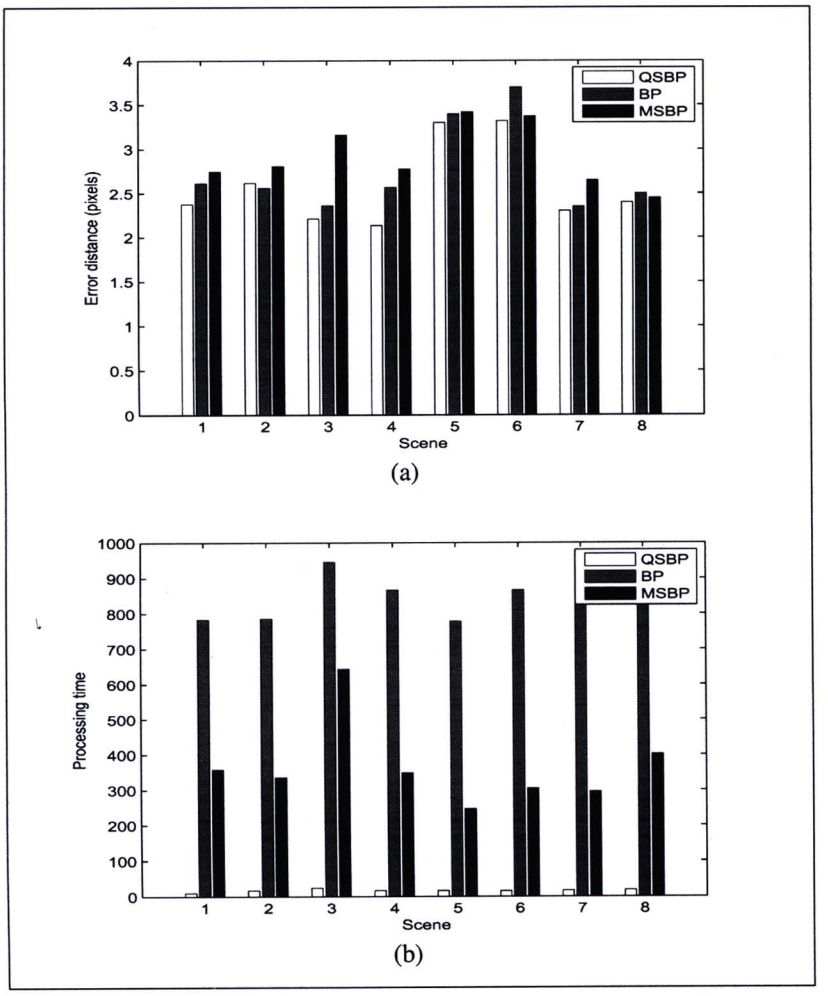

Figure 5.12 A performance comparison between the proposed method and BP and MSBP [29] are shown by white, gray and black bars, respectively (a) Accuracy (b) Efficiency

5.5 Conclusion

We propose a part-based tracking algorithm by integrating Quick Shift, a simple and efficient mode seeking method, into the belief propagation framework. The main idea is to find the mode of marginal probability of belief propagation to be used to predict points in the next iteration and, in that way, only samples around modes are computed in each iteration of belief propagation. Therefore, our proposed method needs fewer samples than NBP or MSBP [29]. In addition, it converges to the best solution faster than the other methods. This approach can reduce the computational complexity dramatically due to the reduction of search space while preserving accuracy. In addition, we apply feedback information of 3D human pose in the first iteration of belief propagation for predicting state. The method was experimented on several videos and the results showed very good performance and robustness in both accuracy and efficiency.

The 2D human joint points of each frame are used for reconstruction and tracking of 3D articulated human pose. The method is explained in Chapter 6