CHAPTER 6 RECONSTRUCTION AND TRACKING OF 3D-ARTICULATED HUMAN BODY FROM 2D POINT CORRESPONDENCES OF A MONOCULAR IMAGE SEQUENCE

In the previous chapter, 2D human joint points of each frame are obtained and then used for reconstruction and tracking of 3D articulated human pose. This chapter presents our method to estimate and track 3D relative positions of an articulated human body from point correspondences of an uncalibrated monocular image sequence. It is based on a geometric constraint of body parts in the skeleton model. Moreover, it applies the concept of Multiple Hypothesis Tracking (MHT) [115] with a motion-smoothness function between consecutive frames to automatically find the optimal solution for this non-uniqueness of solutions. The MHT also helps in recovering from incorrect tracking due to incorrect decision. We also compared the performance of our proposed method with other techniques. Our method outperform other approaches, especially in scenes with strong perspective effects and difficult poses.

This chapter is organized as follows. Section 6.1 begins by surveying previous work of 3D reconstruction using camera-based methods that inspire the work in this chapter. Sections 6.2 and 6.3 describe the proposed articulated body reconstruction and 3D body tracking methods, respectively. Experimental results and evaluations are presented in Section 6.4. Finally, conclusions are drawn in Section 6.5.

6.1 Previous Work

One of the approaches to reconstruct a 3D human pose from the single view group is based on a camera-model concept. This method attempts to lift 2D correspondence data to construct a 3D model of a human pose, normally formed as a skeleton model. The 2D joint correspondence positions that serve as its input are usually obtained by hand labeling [61, 18, 116], matching to a marked image [9], performing 2D human body tracking [24] or detecting from installed markers [117]. 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], that root joint of human body moves parallel to the image plane without Z direction displacement [110] or that camera parameters are known a priori [25].

The construction of a 3D-articulated human pose using a single camera data is a popular approach because it is simple and does not require a database of 3D known human pose or an extensive computation. There are two state-of-the-art approaches in the camera-model-based group, namely, scaled-orthographic and perspective approaches. The scaled-orthographic approach is a well-known technique proposed by Taylor[18] and used widely in many research works [19, 20, 9, 21, 22, 23, 24]. It does not require knowledge of camera parameters. In this approach, the basic assumption is that most human parts is close to a plane parallel to the image plane. The scaled-orthographic

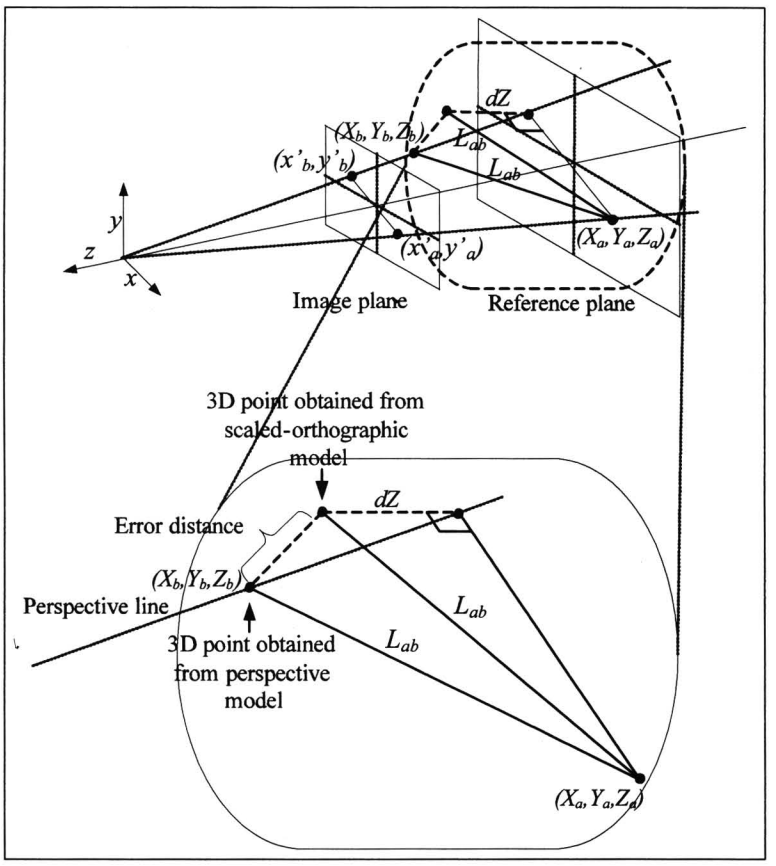

Figure 6.1 The scaled-orthographic camera model of two pairs of corresponding points

model is shown in Figure 6.1. By assuming that the actual length of a segment ab, L_{ab}, is known, we can write

$$L_{ab}{}^2 = (X_a - X_b)^2 + (Y_a - Y_b)^2 + (Z_a - Z_b)^2, \tag{6.1}$$

where (X_a, Y_a, Z_a) and (X_b, Y_b, Z_b) are real-world coordinates of points a and b, respectively. The segment is projected onto the image plane by a scaling factor, s; therefore,

$$\begin{aligned} (x'_a - x'_b) &= s(X_a - X_b) \\ (y'_a - y'_b) &= s(Y_a - Y_b), \end{aligned} \tag{6.2}$$

where (x'_a, y'_a) and (x'_b, y'_b) are two corresponding points, on the image plane, of points a and b, respectively. By assuming that the actual point is projected perpendicularly to the reference plane, the relative depth between the two points is defined as

$$\begin{aligned} dZ &= (Z_a - Z_b) \\ dZ &= \sqrt{L_{ab}{}^2 - \frac{(x'_a - x'_b)^2 + (y'_a - y'_b)^2}{s^2}}. \end{aligned} \tag{6.3}$$

Because dZ cannot be complex,

$$s \geq \frac{\sqrt{(x'_a - x'_b)^2 + (y'_a - y'_b)^2}}{L_{ab}}. \tag{6.4}$$

To obtain the most accurate 3D position from the model, the optimal scaling factor is searched using a comprehensive grid-based optimization. Once the real-world depth of the first point, Z_a, is defined, the depth of the next point, Z_b, is computed using equation (6.1) and equation (6.2). In the reconstruction of 3D articulated human by the concept of this model, the corresponding points on the image plane and the length of each segment are assumed to be known a priori. This method is simple to establish a 3D pose; however, it may not yield accurate results. It is most suitable for telecentric cameras used in most industrial inspection applications, but not for perspective cameras found in general. An example of the error distance of the 3D reconstruction from the scaled-orthographic model is shown in Figure 6.1. It is significantly increased with the perspective effects (i.e., in lower magnification systems).

To overcome the limitations of the scaled-orthographic approach, Remondino et al. propose a method based on a perspective concept [25]. In their work, some camera parameters such as focal length must be known a priori. Moreover, at least one predefined segment is assumed to lie on a plane parallel to the image plane. By assuming that the scaling factor is

$$s = \frac{F}{Z} \tag{6.5}$$

where F is the principal distance (approximated by the focal length of the lens). The relationships of points on the real-world and the image plane can be represented by

$$\begin{aligned} s_a X_a &= x'_a & \qquad s_b X_b &= x'_b \\ s_a Y_a &= y'_a & \qquad s_b Y_b &= y'_b, \end{aligned} \tag{6.6}$$

where (X_a, Y_a, Z_a) and (X_b, Y_b, Z_b) are real-world coordinates of points a and b and are projected onto the image plane at points (x'_a, y'_a) and (x'_b, y'_b) by the scaling factors s_a and s_b, respectively. L_{ab} is the known actual length of the segment ab. Similar to equation (6.1),

$${L_{ab}}^2 = (\frac{x'_a Z_a - x'_b Z_b}{F})^2 + (\frac{y'_a Z_a - y'_b Z_b}{F})^2 + (Z_a - Z_b)^2. \tag{6.7}$$

From the assumption that F is known and at each frame at least one predefined segment lies on a plane parallel to the image plane ($Z_a = Z_b$), Z_b value of that segment can be solved by

$$Z_b = \frac{F L_{ab}}{\sqrt{{x'_a}^2 - 2x'_a x'_b + {x'_b}^2 + {y'_a}^2 - 2y'_a y'_b + {y'_b}^2}}. \tag{6.8}$$

After the depth of the joints on the image plane is computed by equation (6.8), the depth of the next point of the next connecting segment can be computed using equation (6.7). By the concept of this model, the focal length, the corresponding points on the image plane and the length of each segment are assumed to be known a priori. Moreover, this model needs a basic assumption that at least one predefined segment lies parallel to the image plane.

In general, this approach is more accurate than the scaled-orthographic approach. However, it is still restricted by the assumption that at least one predefined segment lies on a plane parallel to the image plane. This is difficult to meet in most general

image sequences.

In reconstruction of 3D human pose from 2D corresponding point approach, such 2D joint inputs are usually based on pixel coordinate system of an image that pixel $(0, 0)$ lies on left-top of the input image. By reconstruction of 3D human pose based on a camera-concept model, coordinate $(0, 0)$ is referred to center of the image. Then, such 2D joint inputs are converted to the image coordinate system before reconstruction of 3D human pose. Figure 6.2 shows relationship of real-word coordinate system, reference coordinate system and image coordinate system. The 3D real-word coordinate system is actual coordinate of human, while the 3D reference coordinate system is 3D coordinate in a virtual distance (as used in our approach). The 3D human pose at the reference coordinate system is a rescale of that in the real-world coordinate system.

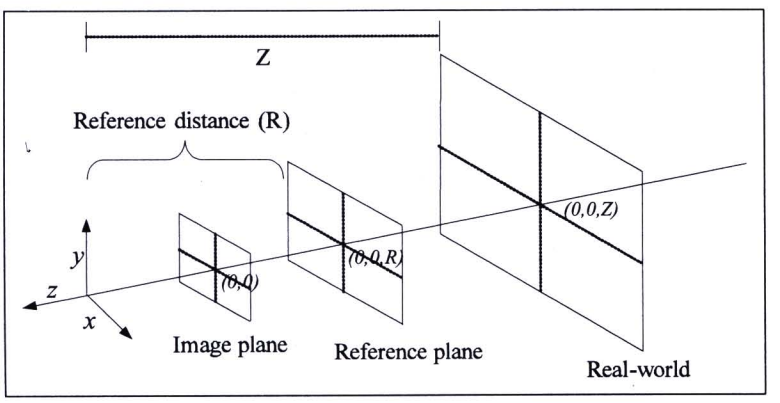

Figure 6.2 The real-word coordinate system, reference coordinate system and image coordinate system

6.2 3D Articulated Body Reconstruction

In our approach, the construction of the 3D relative body is performed from known 2D correspondences based on a 3D skeleton model. Since the relative 3D pose can be constructed at any scale, we therefore define a reference distance as a virtual distance which the root-node joint lies in the first frame. Firstly, the relative length of each segment is determined and then the reference distance is computed for the first frame. These parameters are used in the reconstruction of the 3D body in the subsequent frames.

The 3D human skeleton in our work is modeled by 15 joint points and 14 segments. 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. The skeleton model can be expressed by a tree structure with a root node. Note that in our work, the abdomen joint is selected as the root, since it is relatively more stable than other joints. The tree structure of the human model with the abdomen root node has level of tree less than that with other joints as the root because the abdomen joint is central joint of the tree structure of the human model. It leads to less error accumulation in the joint position computation than the other joints. Moreover, the abdomen has relatively small movements than the other joints of human body. We

also define a relative length of each part based on the concept of Leonardo Da Vinci [25], i.e. each part is up to a scale T, as shown in Table 6.1. However, it is sometimes difficulty to fit human model by the concept of Leonardo Da Vinci [25]. To alleviate this problem, we can consider corresponding symmetric parts and introduce relaxation into the reconstruction.

Table 6.1 The relative length of each segment in the human model

Body parts	Relative length
Head	1.25 T
Torso	2.50 T
Shoulder	2.00 T
Upper arm	1.50 T
Lower arm	2.00 T
Pelvic	1.50 T
Upper leg	2.00 T
Lower leg	2.00 T

6.2.1 Problem Formulation

Our approach for the 3D reconstruction of a relative human pose from known coordinates of 2D-point correspondence is based on a perspective camera concept, as shown in Figure 6.3. The 2D points obtained from an uncalibrated camera serve as input for our work.

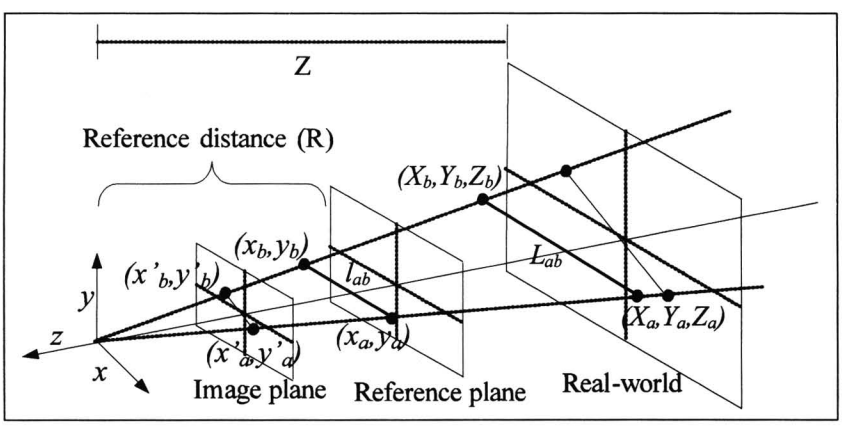

Figure 6.3 The perspective camera model of two pairs of corresponding points

In Figure 6.3, $V_a = (X_a, Y_a, Z_a)$ and $V_b = (X_b, Y_b, Z_b)$ are real-world coordinates of points a and b, respectively. $v_a = (x_a, y_a, z_a)$ and $v_b = (x_b, y_b, z_b)$ are corresponding relative coordinates of the points a and b around the reference distance, respectively. By the word "around the reference distance", we mean both 3D points are scaled using the same magnification $(\frac{R}{Z})$ where Z is the real world depth of the root node joint and R is the reference distance from the origin to the reference plane and are related to the corresponding points in the reference plane by a scale factor. (x'_a, y'_a) and (x'_b, y'_b) are the respective image coordinates of points a and b at the image plane. L_{ab} and l_{ab} are the lengths between points a and b in terms of real-world length and virtual length

around the reference distance, respectively. The magnification of a coordinate around the reference distance to the real-world coordinate is represented by

$$\frac{1}{\alpha} = \frac{x_a}{X_a} = \frac{y_a}{Y_a} = \frac{z_a}{Z_a} = \frac{l_{ab}}{L_{ab}} = \frac{R}{Z}, \tag{6.9}$$

From Figure 6.3, we can project the points $v_a = (x_a, y_a, z_a)$ and $v_b = (x_b, y_b, z_b)$ onto the image plane and obtain the following relationships:

$$\begin{aligned} x_a &= s_a x'_a & x_b &= s_b x'_b \\ y_a &= s_a y'_a & y_b &= s_b y'_b \\ z_a &= s_a R & z_b &= s_b R, \end{aligned} \tag{6.10}$$

where s_a and s_b are scaling factors corresponding to the ratios of points Z_a and Z_b to R, respectively. To estimate 3D positions from the known 2D point correspondences, we find that

$$\begin{aligned} X_a &= \alpha s_a x'_a & X_b &= \alpha s_b x'_b \\ Y_a &= \alpha s_a y'_a & Y_b &= \alpha s_b y'_b \\ Z_a &= \alpha s_a R & Z_b &= \alpha s_b R. \end{aligned} \tag{6.11}$$

From the obtained parameters, we can recover the 3D relative coordinates of each node around the reference distance using equation (6.10). These are up-to-scale coordinates of the corresponding 3D real-world coordinates. To estimate 3D real-world coordinates, knowledge of either a 3D real-world coordinate or an actual edge length is required by equation (6.11). However, up-to-scale coordinates are adequate for most character-pose-related applications.

6.2.2 Determining Reference Distance

In our approach, the reference distance is required and is determined in the first frame. We set an assumption that a set of three known joints forming a right triangle with only one leg lain on a plane parallel to the image plane in the first frame. This can be easily achieved by a human stands vertically parallel to the image plane and not all of his/her joints lie on a plane parallel to the image plane. An example of the human pose satisfying the assumption is shown in Figure 6.4. First, we find the reference distance by considering a segment that is not in a plane parallel to the image plane and denoted it by ε_{qr} with length l_{qr}, as shown in Figure 6.4. For segment selection, a segment with right angle with respect to the vertical axis of the human body is chosen. In Figure 6.4, $v_p = (x_p, y_p, z_p)$, $v_q = (x_q, y_q, z_q)$ and $v_r = (x_r, y_r, z_r)$ are virtual coordinates of segment end points around the reference distance. l_{pq} is parallel to the image plane and can be determined using Euclidean distance of its endpoint image coordinates. l_{qr} is proportional to l_{pq} according to the Da Vinci's ratio. For the reference distance computation, the projected end points of the link are used to obtain

$${l_{qr}}^2 = (s_q x'_q - s_r x'_r)^2 + (s_q y'_q - s_r y'_r)^2 + (s_q R - s_r R)^2.$$

Hence, the reference distance can be described by the following equation:

$$R = \sqrt{\frac{{l_{qr}}^2 - (s_q x'_q - s_r x'_r)^2 - (s_q y'_q - s_r y'_r)^2}{(s_q - s_r)^2}}, \tag{6.12}$$

where (x'_q, y'_q) and (x'_r, y'_r) are image coordinates of points q and r, respectively; s_q and s_r are the respective corresponding scaling factors of points q and r; and l_{qr} is its length.

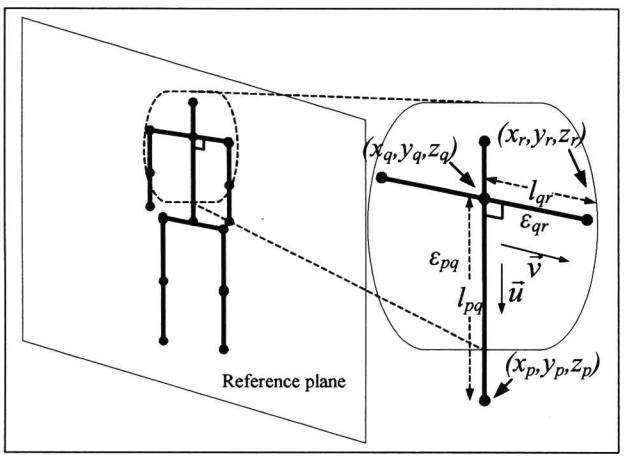

Figure 6.4 Initial posture in the first frame that two joints are not in the same plane

From equation (6.12), s_q and s_r must be calculated in the determination of the reference distance. From the assumption made in the first frame, the axis of the scaled human body is set to lie on the reference plane such that scaling factors regarding the abdomen and neck joints are set to unity. In the determination of the reference distance, we select the link between the neck and the abdomen joints as vector $\vec{u}$ and the link between the neck and the left shoulder joints as vector $\vec{v}$. The scaling factor of the abdomen and neck joints is 1 and $\vec{u}$ is perpendicular to $\vec{v}$, i.e., $\vec{u}.\vec{v} = 0$ (only required for the first frame), and hence

$$\begin{aligned} \vec{u} &= (x'_p - x'_q, y'_p - y'_q, 0) \\ \vec{v} &= (s_r x'_r - x'_q, s_r y'_r - y'_q, s_r R - R) \end{aligned}$$

and

$$s_r = \frac{x'_p x'_q - {x'_q}^2 + y'_p y'_q - {y'_q}^2}{x'_p x'_r - x'_q x'_r + y'_p y'_r - y'_q y'_r}. \tag{6.13}$$

(x_p, y_p, z_p), (x_q, y_q, z_q) and (x_r, y_r, z_r) are the abdomen, neck and left shoulder joints around the reference distance and are projected onto the image plane at points (x'_p, y'_p), (x'_q, y'_q) and (x'_r, y'_r) by scaling factor s_p, s_q and s_r, respectively. By putting $s_q = 1$ and s_r obtained from equation (6.13) back to Eq. (6.12), the reference distance is determined.

6.2.3 Determining Scaling Factor of Root-node

The position of the root segment is very important, because it is the starting point for the reconstruction of the 3D body in our approach. From our assumption, the scaling factor of the root-node joint, s_1 is unity at the first frame. However, it changes and, therefore, requires computation for subsequent frames due to human motion. It is determined by a comprehensive grid-based search with the constraint that its value

cannot be negative. We can also perform a faster search for s_1 within a small range around s_1 of the previous frame. The values satisfying 3D human pose based on a joint-angle limitation constraint (Section 6.3.1)) are selected as s_1 in the current frame.

6.2.4 Determining Scaling Factor of other Nodes

For the reconstruction and track 3D relative positions of articulated body, the scaling factors of each segment is computed. Suppose s_a is known from result of the previous computed segment, s_b can be determined by finding intersection points between a sphere and a line. The starting joint $(s_a x'_a, s_a y'_a, s_a R)$ and length l_{ab} of segment ε_{ab} served as the center and the radius of the sphere, respectively. By this assumption that, we can write

$$(x_b - s_a x'_a)^2 + (y_b - s_a y'_a)^2 + (z_b - s_a R)^2 = {l_{ab}}^2. \tag{6.14}$$

The line started from the origin $(0, 0, 0)$ pointing in the direction (x'_b, y'_b, R) meets the sphere at point (x_b, y_b, z_b) as shown in Figure 6.5 and we have

$$\begin{aligned} x_b &= s_b x'_b \\ y_b &= s_b y'_b \\ z_b &= s_b R. \end{aligned} \tag{6.15}$$

Since (x'_a, y'_a) and (x'_b, y'_b) are points on the image plane, R and l_{ab} are computed from the first frame, only s_b remains to be calculated. We also perform the determination for $s_k, k = 2, ..., N$ sequentially using s_a obtained from the result of the previous computed segment. By replacing equation (6.15) back to Eq. (6.14), we arrive at

$$\begin{aligned} ({x'_b}^2 + {y'_b}^2 + R^2){s_b}^2 - (2s_a x'_a x'_b + 2s_a y'_a y'_b + 2s_a R^2)s_b \\ +({s_a}^2 {x'_a}^2 + {s_a}^2 {y'_a}^2 + {s_a}^2 R^2 - {l_{ab}}^2) = 0. \end{aligned}$$

Therefore,

$$s_b = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}, \tag{6.16}$$

where

$$\begin{aligned} A &= {x'_b}^2 + {y'_b}^2 + R^2 \\ B &= -(2s_a x'_a x'_b + 2s_a y'_a y'_b + 2s_a R^2) \\ C &= {s_a}^2 {x'_a}^2 + {s_a}^2 {y'_a}^2 + {s_a}^2 R^2 - {l_{ab}}^2. \end{aligned}$$

In our approach, the scaling factor can not be complex or negative. The solutions of scaling factor from equation (6.16) are shown by dot markers in Figure 6.6 (a) and (b). However, it can be seen that the solutions are normally not unique as shown in Figure 6.6 (b). This problem is inherent to all lift-up approaches. To select the best configuration, some further assumptions, e.g. that all joints lie on the same plane [18], are normally assumed. For our technique, we delay the selection of each best configuration of the first two frames to the third frame and use MHT algorithm for the selection process. This makes our method even less restricted.

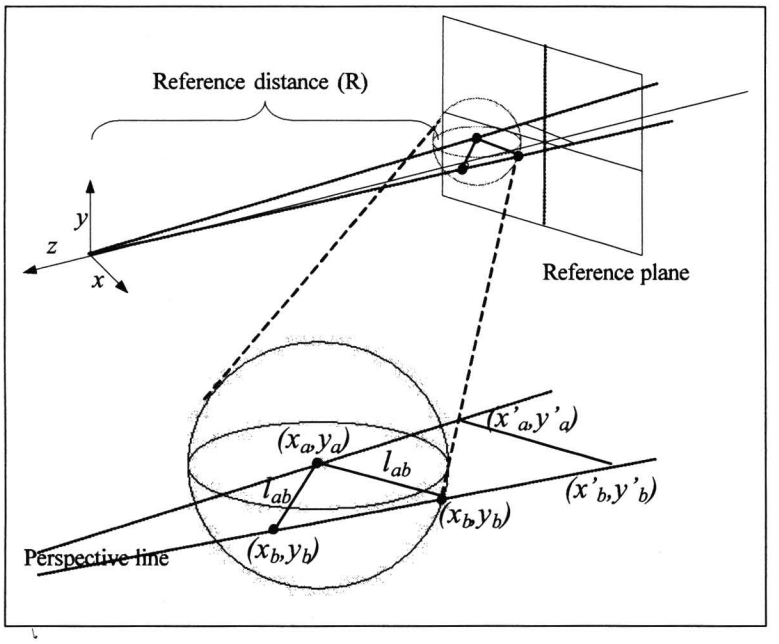

Figure 6.5 The intersection points between a sphere and a line

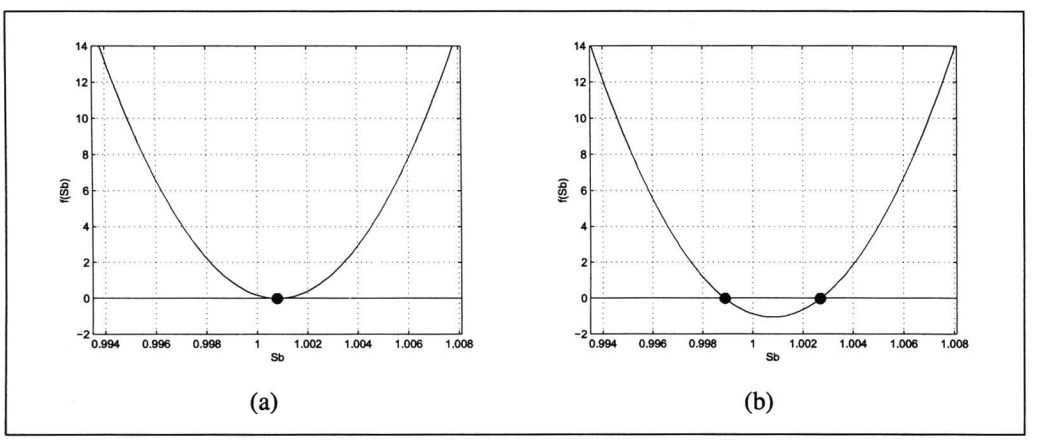

Figure 6.6 Scaling factor (a) one solution (b) two solutions

6.3 3D Human Body Tracking

After the first frame, 3D reconstruction can be more conveniently performed using tracking. At each frame, multiple configurations of 3D reconstructions are obtained. To find the optimal solution, we applied a joint angle limitation constraint in order to prune spurious solutions and to avoid singularity. Moreover, we employed a technique based on the concept of multiple hypothesis tracking (MHT) [115] with a motion-smoothness function to select the best solution based on temporal information.

6.3.1 The Joint Angle Limitation Constraint

Multiple configurations per frame are obtained due to non-uniqueness of the reconstruction. Some configurations are singular or spurious solutions. To remove some of these undesired configurations, we employ a joint angle limitation that limits arm and leg angles to fall between 10 and 180 degrees, as shown in Figure 6.7.

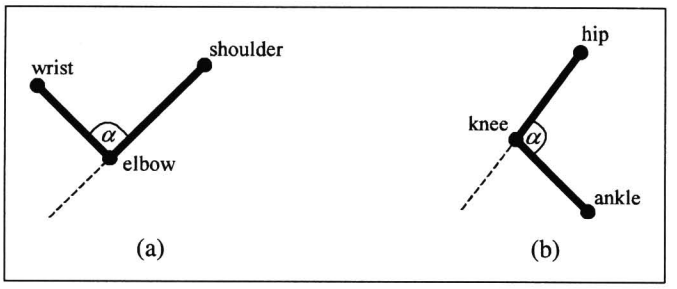

Figure 6.7 Joint angle limitation (a) Arm (b) Leg

6.3.2 Optimal Solution Selection by MHT

Even with the constraint specified, there are still a great number of possible configurations left. Techniques using temporal information have been introduced to select the optimal solution in other works. Most of these approaches are based on choosing the closest configuration from one frame to the next [21, 25, 116]. This method only maintains the solution for the last frame. It works well in some simple poses but fails in a larger number of cases such as change direction immediately of human body; it can also lead to divergences due to an incorrect decision in early frames, especially in arbitrarily complicated poses. One obvious solution is to maintain all possible pose sequences and selects the best solution according to a suitable criterion [10]. However, this makes inhibitive computations as number of frames grows. We therefore proposed an efficient technique based on MHT to select the best configuration at the current frame using the past temporal information while still maintaining a number of most-likely previous poses.

We define the *temporal joint trajectory* of a joint as a time sequence of the distances from the 3D position of that particular joint to the coordinate origin, i.e., the camera center. The smoothness of the trajectory is used to evaluate the fitness of a hypothesis instead of the hypothesis probability used in the original study [115]. A polynomial function and least squares method can be used to develop a motion model as well as an estimator in the context of MHT. In our experiment (as shown in Section 6.4.3), we found that a simple line connecting two previous frames of the joint sequence was adequate and very efficient, and therefore, was used it in our work. The details on the formulation of the technique are described below.

At frame t, a pose configuration is denoted by $\theta_i(t)$, and the set of all possible configurations is represented by $\{\theta_i(t)\}$, $i = 1, 2, ..., N_t$. From MHT concept, a hypothesis at frame t, which is denoted by Θ_i^t, is a pose sequence and can be written as

$$\Theta_i^t = \{\Theta_{m(i)}^{t-1}, \theta_i(t)\}, \tag{6.17}$$

where $\Theta_{m(i)}^{t-1}$ is the parent pose sequence at frame $t-1$ of Θ_i^t. To determine the fitness of a hypothesis, the smoothness of the trajectory is used and computed by

$$E_i^t = \Delta(\theta_i(t)) + E_{m(i)}^{t-1}, \tag{6.18}$$

where

$$\Delta(\theta_i(t)) = \sum_{j=1}^{N} \frac{|t + A_{i,j}^t D_{i,j}^t + B_{i,j}^t|}{\sqrt{A_{i,j}^t{}^2 + 1}}. \tag{6.19}$$

E_i^t is the fitness of the pose sequence i at frame t, and $D_{i,j}^t$ is the distance from joint j of the configuration i at frame t to the origin. $A_{i,j}^t$ and $B_{i,j}^t$ are parameters of the line equation passing through the points $(t-2, D_{i,j}^{t-2})$ and $(t-1, D_{i,j}^{t-1})$, respectively. N is the number of joints in the human body model.

The fitness value is calculated for each of the possible pose sequences in each hypothesis of MHT. The fitness values in each hypothesis of MHT are then ordered. The configuration corresponding to the best fitness is selected as the solution for that frame, while k best configurations according to their fitness values are kept for generating new hypotheses of MHT for the next frame. The computational complexity is $O(2^N k \log(2^N k))$ for N joint body model. Figure 6.8 shows the flowchart of the algorithm.

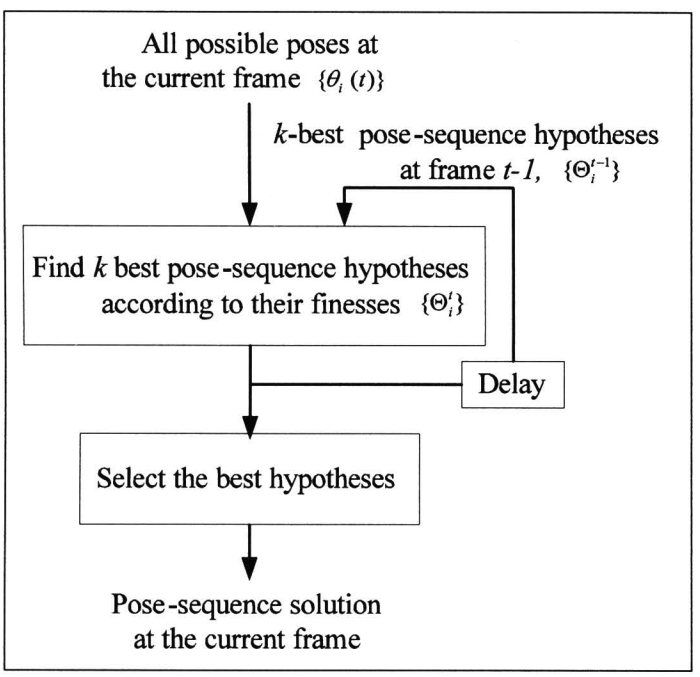

Figure 6.8 Flow chart of the multiple hypothesis algorithm

Our method bears similarity to Howe's [10] work in the sense that more temporal information than the previous frame is used. However, Howe's method performs a comprehensive search of all possible pose sequences in the past few frames, while our method requires a search of only the k best possible pose sequences in the past few frames. In addition, Howe's 3D position configurations per frame are obtained by consulting an image feature database; thus, the number of configurations obtained may be different from ours.

6.4 Experiments and Evaluations

To evaluate the performance of our proposed method, we tested our approach on both synthesized and real-world image sequences. We assume that 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. The 2D joint correspondence positions that served as input are usually obtained by hand labelling or by 2D human body tracking as described in previous chapter. To benchmark the performance of our proposed method, we compare our approach with both scaled-orthographic [18] and previous perspective approaches [25]. Moreover, we tested it on several real-world image sequences. In all of our experiments, we selected the link between the neck and left shoulder of the human for the reference distance computation due to its perpendicular angle assumption as well as the abdomen joint for the root joint due to its relatively small movements. To evaluate the performance of reconstruction of 3D articulated body, 3D relative pose results are scaled to best match and then compared with 3D real world coordinates of the ground-truth. In this experiment, the 1000 best hypotheses are maintained at each frame by our tracking method.

6.4.1 Performance Comparison with other Approaches

We synthesized four 3D human image sequences in order to compare our proposed method with both scaled-orthographic [18] and previous perspective [25] approaches in terms of accuracy. The 2D projected joints from those 3D synthesized data served as the input in the comparison. The image sequences include aerobic style activities, which contain 15 frames in each scene. Attributes of each scene are shown in Table 6.2. Since the orthographic method is originally designed for 3D reconstruction from a single image, it is not fair to perform benchmarking based on tracking results due to the accumulated errors in the latter frames of the sequences. Therefore, we exclude solving an ill-posed problem by selecting the closet configuration to the ground truth for all candidate algorithms at each frame.

Table 6.2 Attributes of four test image sequences

Scene	Magnitude of Perspective Effect	At Least One Segment Lies on a Parallel Plane
Scene 1	Weak	Yes
Scene 2	Weak	No
Scene 3	Strong	Yes
Scene 4	Strong	No

Some ground truth frames of scenes 1-4 are shown in rows 1-4 of Figure 6.9 (a). Examples of results from our approach using scenes 1-4 are shown in rows 1-4 of Figure 6.9 (b). Some results of the scaled-orthographic approach for scenes 1-4 are shown in Figure 6.10 (a). Similar results from the previous perspective technique are presented in Figure 6.10 (b). From the experimental results, the averaged per-joint error distances from the ground truth for our proposed method, the scaled-orthographic approach [18] and the previous perspective [25] approach are shown in Table 6.3. It can be seen that the results of our approach are more accurate than other approaches for all sequences. The scaled-orthographic approach [18] performed very well in the first sequence because the scene contained a weak perspective effect and relatively simple

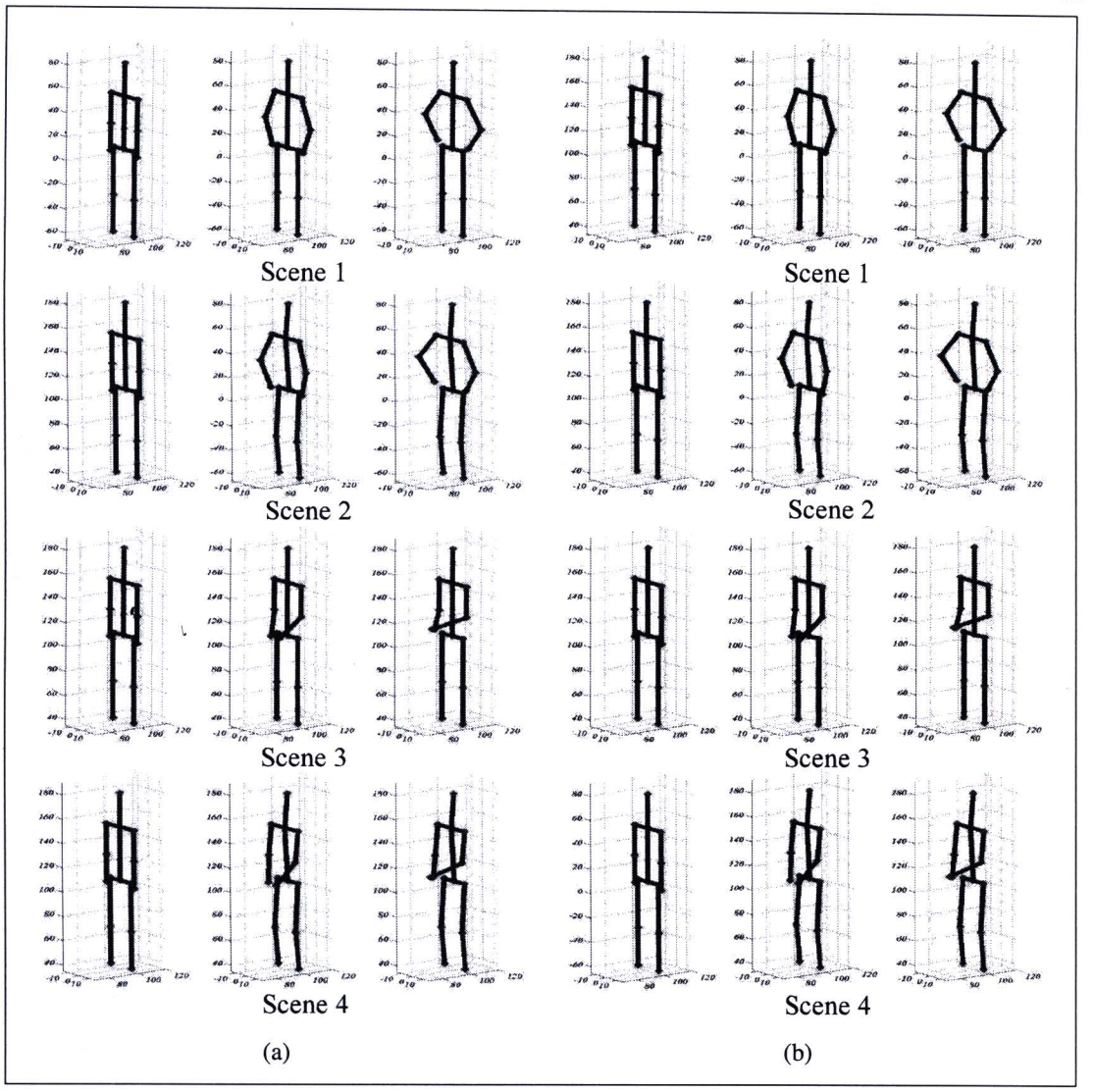

Figure 6.9 Samples of ground truth and results from our approach in synthesized data (a) samples of ground truth for scenes 1-4 and (b) samples of results from our approach

poses. For the previous perspective approach [25], the results were very good and comparable to ours in scenes 1 and 3; this is because each frame of those sequences has at least one segment in a parallel plane. Note that our approach outperforms the others for sequences with both weak and strong perspective effects.

Table 6.3 Averaged per-joint error distances (in centimeters) from different approaches

Scene	Our Approach	Scaled-orthographic Approach [18]	Previous Perspective Approach [25]
Scene 1	0.00	1.60	0.08
Scene 2	0.10	2.34	7.21
Scene 3	0.90	4.25	1.01
Scene 4	0.89	4.92	9.68

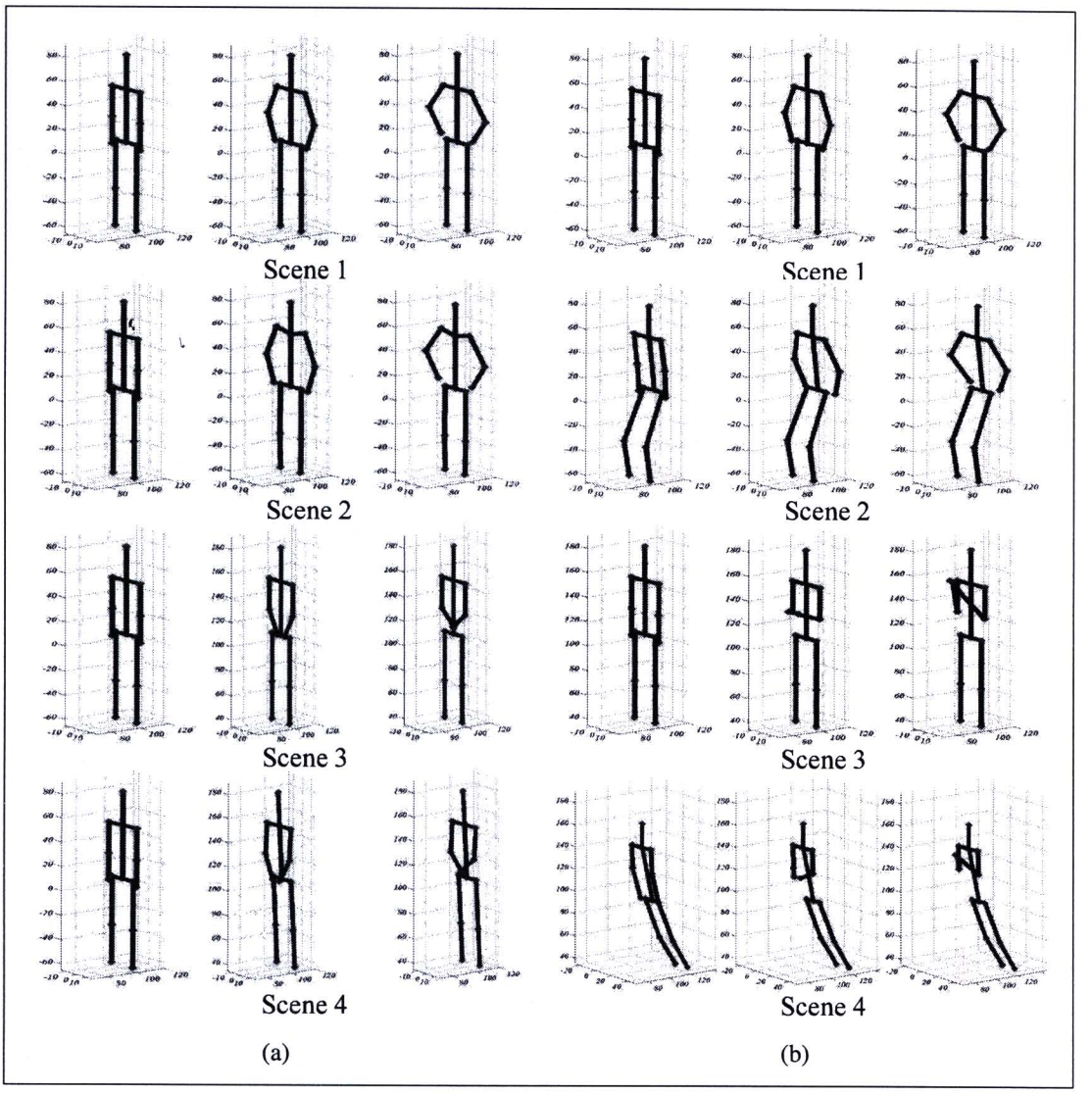

Figure 6.10 Samples of results from Taylor'method and Remondino and Roditakis'method (a) samples of Taylor [18] result for scenes 1-4 and (b) samples of results from Remondino and Roditakis technique [25]

6.4.2 Results on Real-world Image Sequences

In addition, we tested our approach using three real-world image sequences. We used the known 2D position of each joint obtained manually in the image as input data for our proposed method. The real-word image sequences include walking, ball-throwing and back-flipping in a Mocap dataset [119] containing 300, 800 and 239 frames, respectively. Some results of the walking, ball throwing and back flipping scenes from our approach are shown in Figures 6.11, 6.12 and 6.13, respectively.

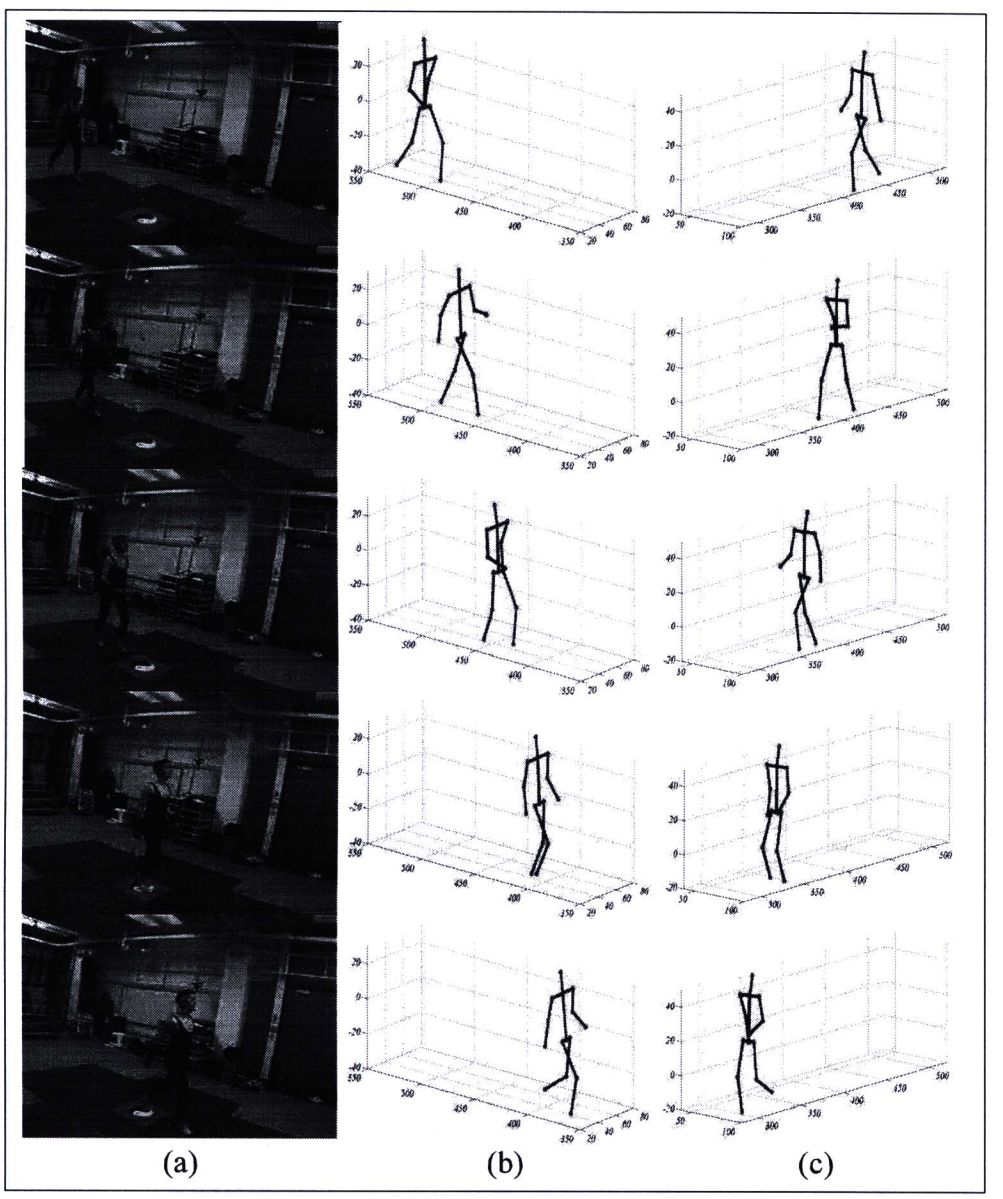

Figure 6.11 Some results for the walking scene (a) original images (b) the same view as the image and c) corresponding results in a different view

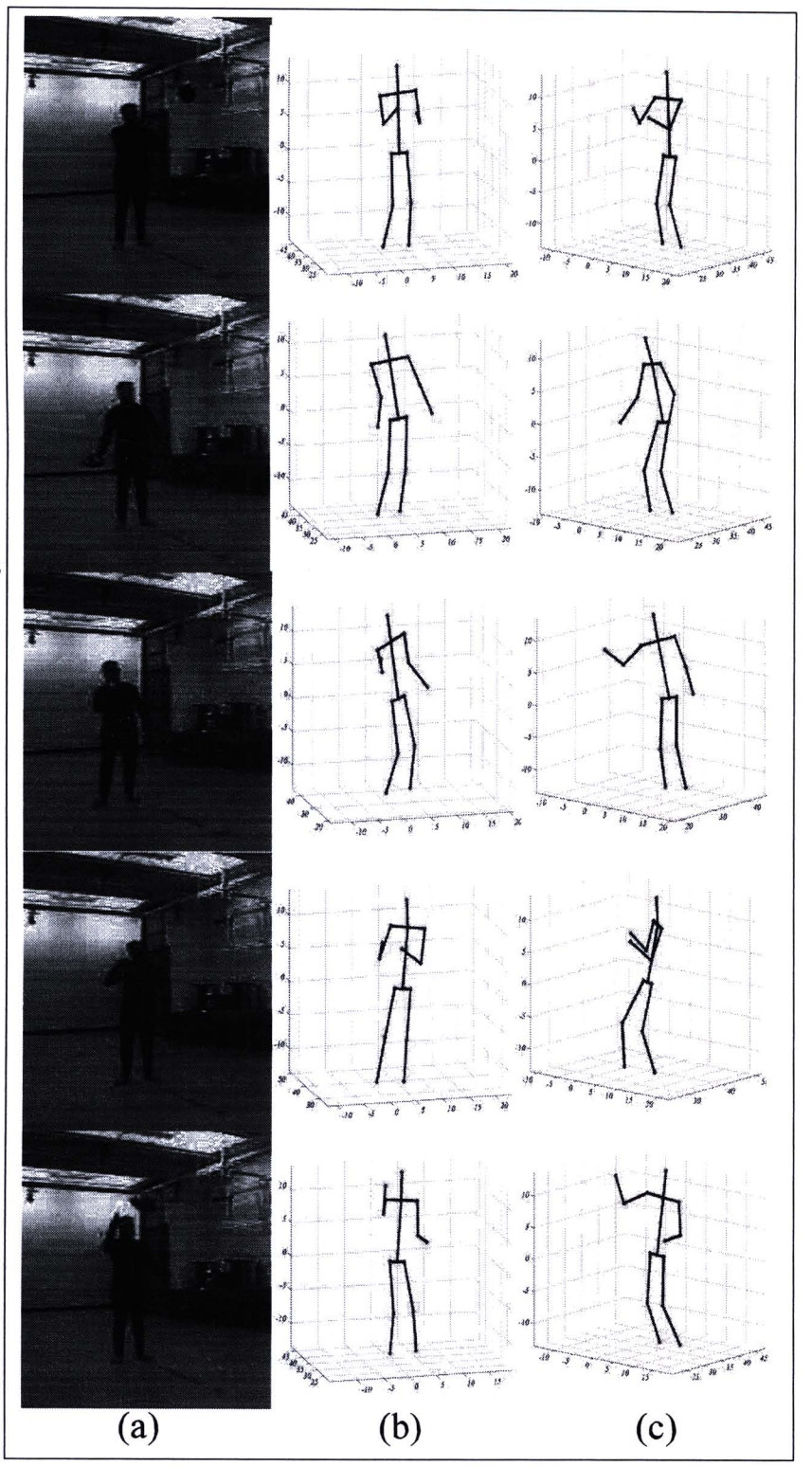

Figure 6.12 Some results from the ball throwing scene (a) original images (b) the same view as the image, and c) corresponding results in a different view

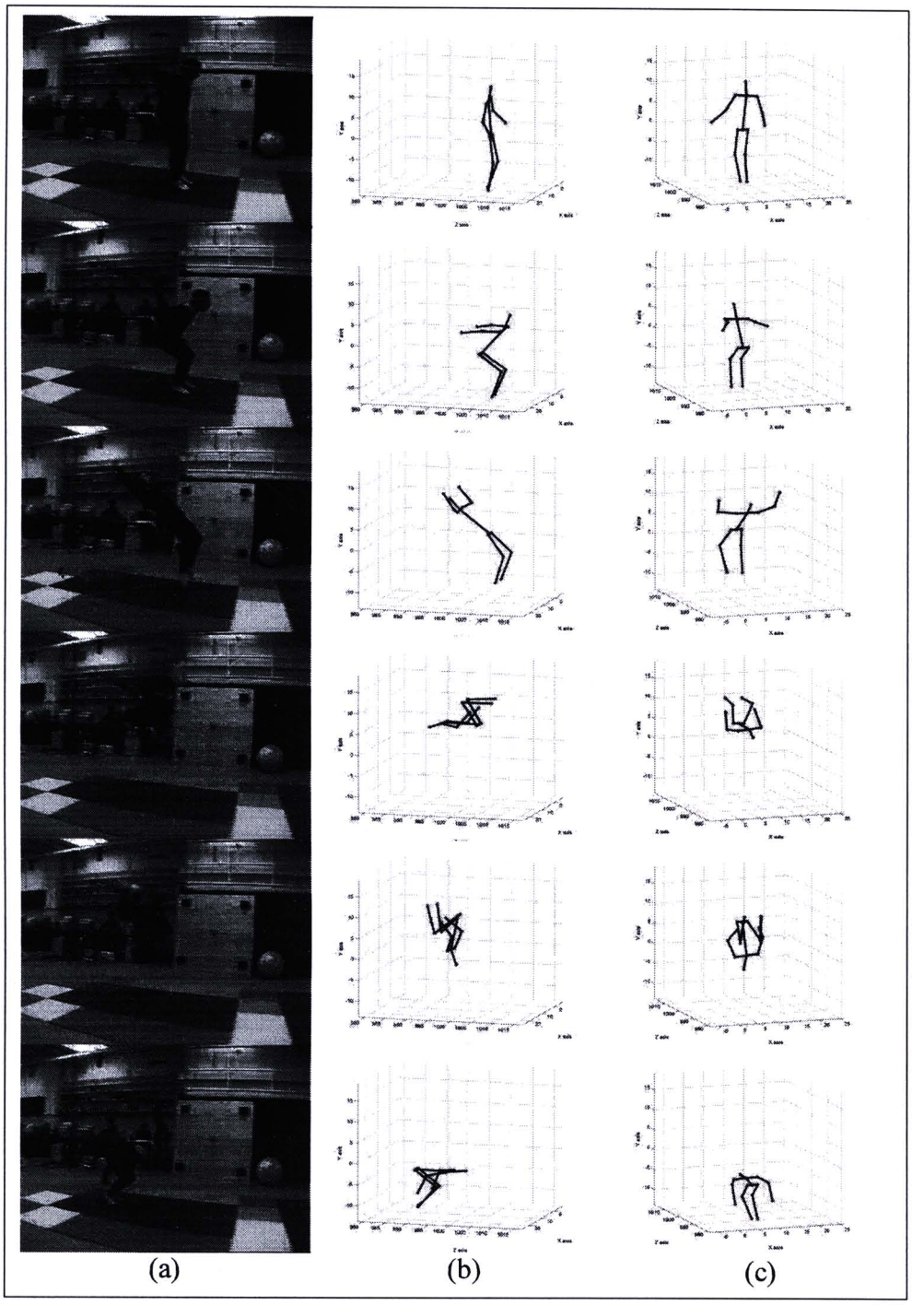

Figure 6.13 Some results on the back flip scene (a) original images (b) the same view as the image and c) corresponding results in a different view

The results show that the accuracy of our proposed method is very high in all sequences. On average, the error distance of the walking, ball-throwing and back-flipping scenes as compared with the corresponding ground truth is 1.25, 1.21 and 1.97 centimeters, respectively. The error distance average of each joint is shown in Figure 6.14. It can be seen that the error distances of elbow and wrist joints are greater than those of other parts. This is because they are moving fast and not smoothly.

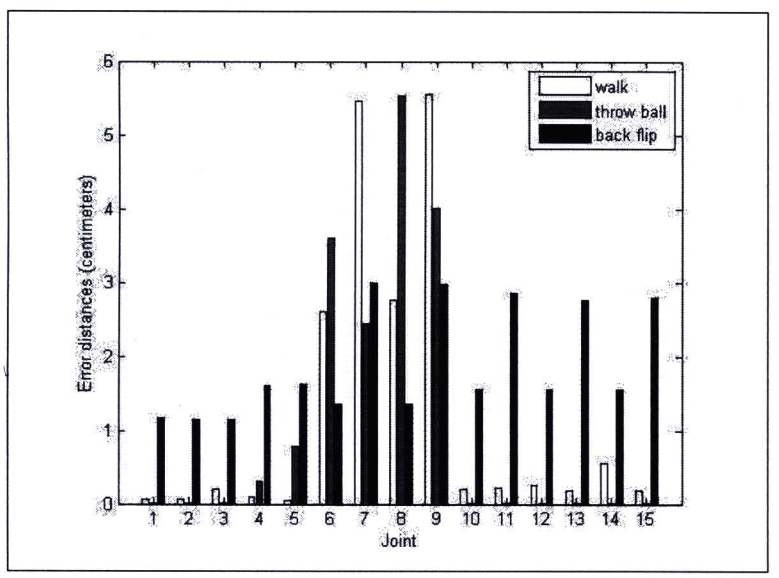

Figure 6.14 The error distances of walking, throwing and back-flipping are shown by white, gray and black bars, respectively

Moreover, we used the 2D positions of each joint obtained from 2D tracking approach presented in the previous chapter as input data. The real-word image sequences include two scenes of aerobics style activities containing 200 and 195 frames, respectively. Moreover, we tested our approach with walking image sequence. Some results from our approach with two scenes of aerobics style activities and walking image sequence are shown in Figures 6.15, 6.16 and 6.17, respectively. The results show that the accuracy of our proposed method is very high in all sequences.

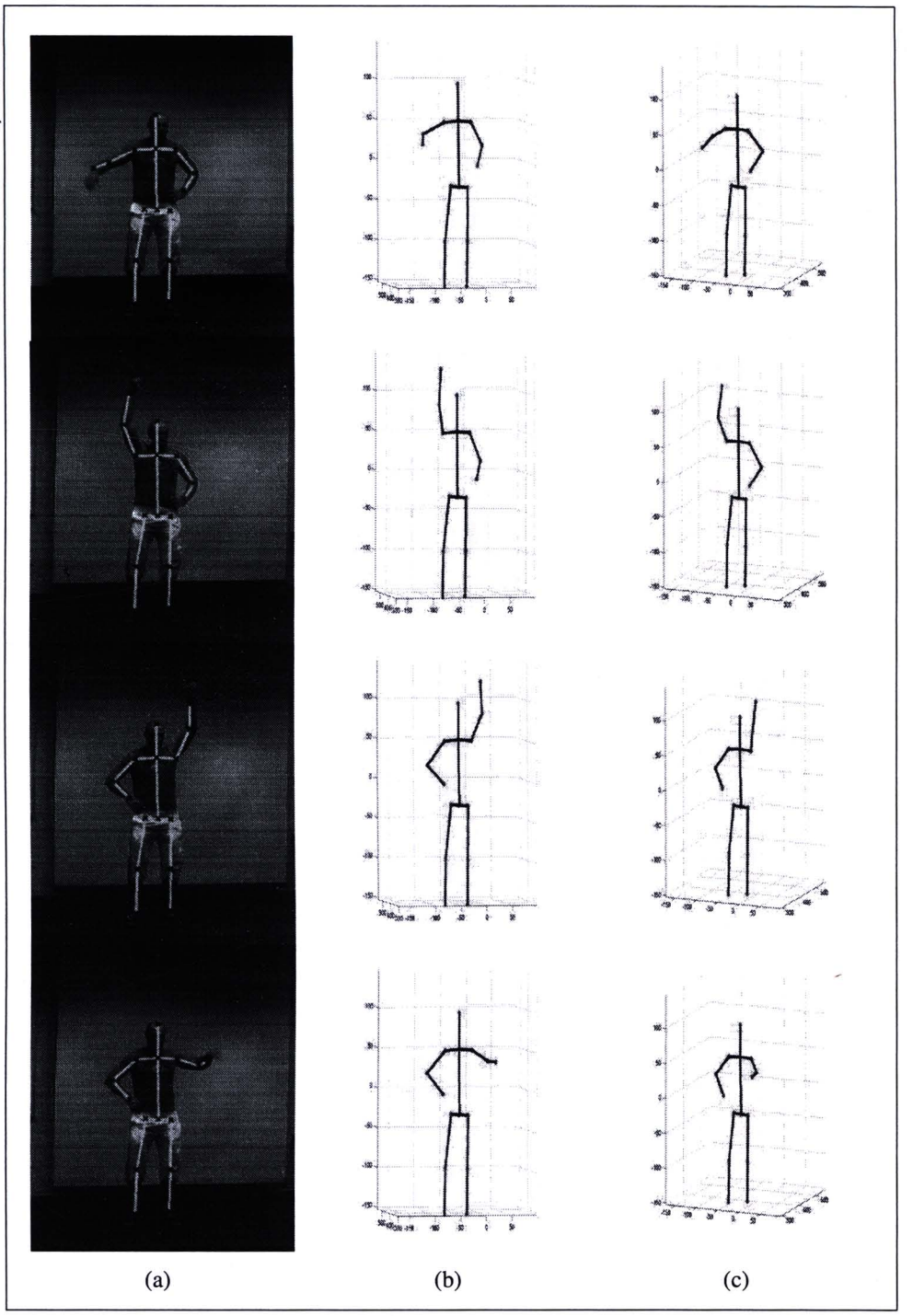

Figure 6.15 Some results on the first image sequence of aerobics style activities (a) original images (b) the same view as the image and c) corresponding results in a different view

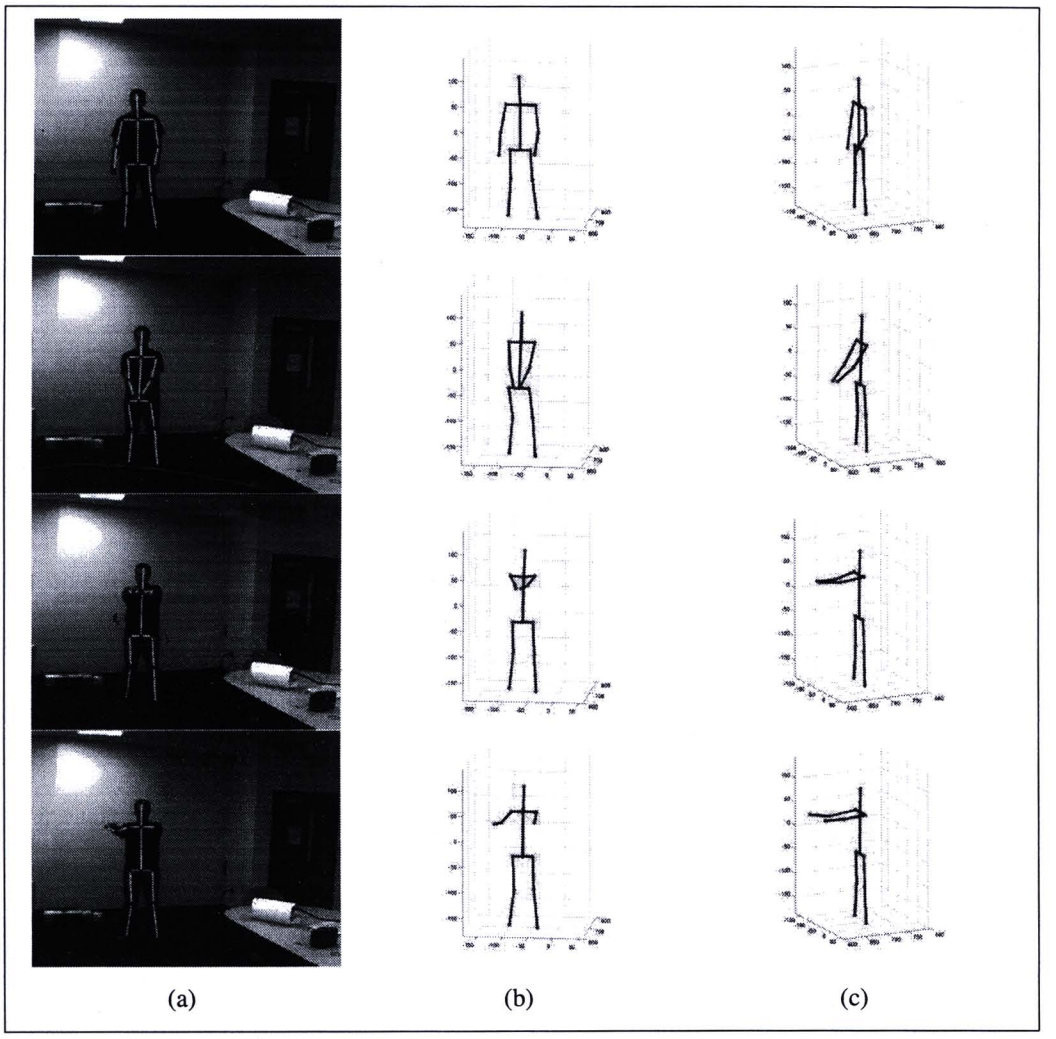

Figure 6.16 Some results on the second image sequence of aerobics style activities (a) original images (b) the same view as the image and c) corresponding results in a different view

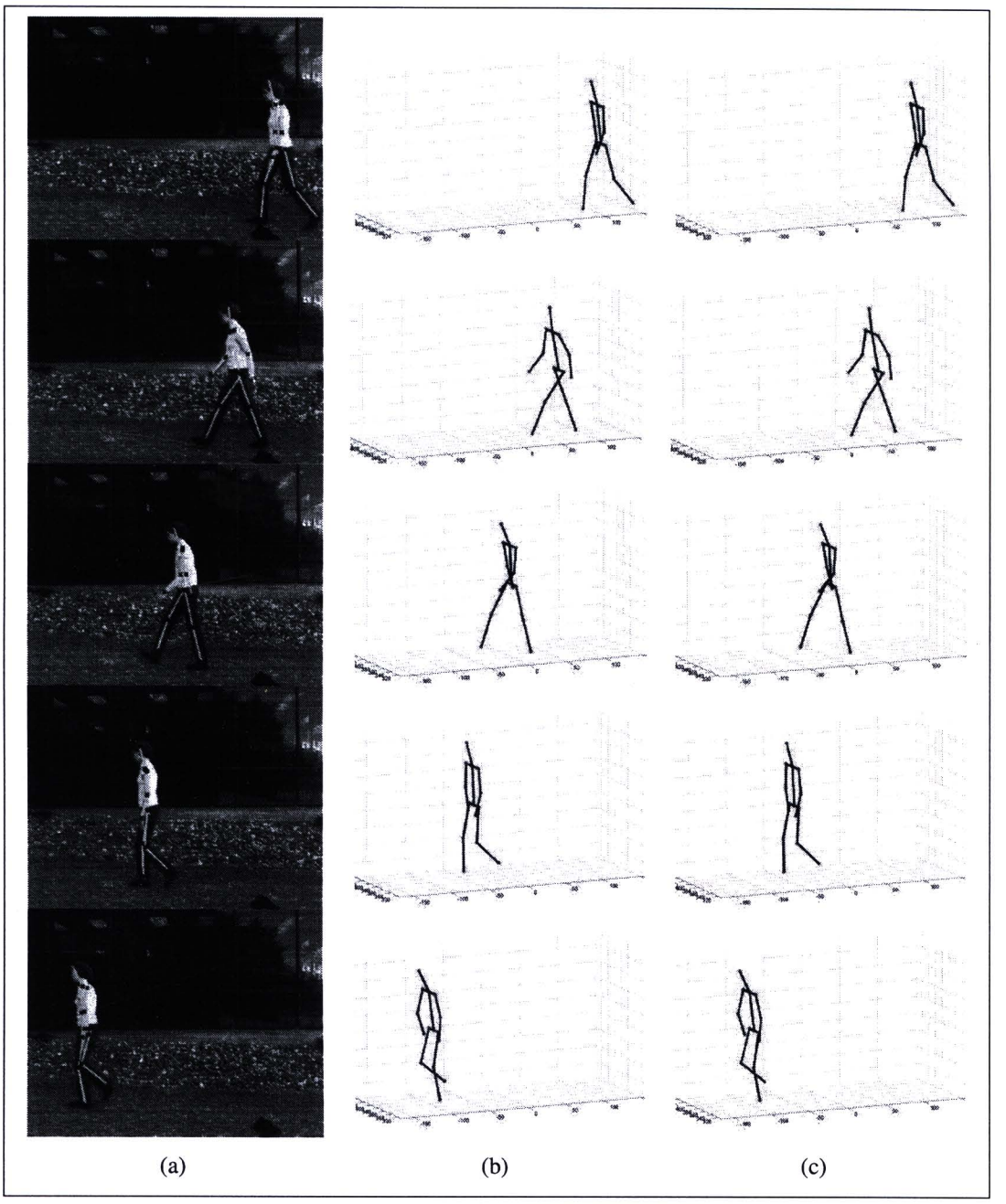

Figure 6.17 Some results on walking image sequence (a) original images (b) the same view as the image and c) corresponding results in a different view

6.4.3 Performance of 3D Tracking with Motion-smoothness Function

To select the best configuration, we therefore proposed an efficient technique based on MHT [115]. The fitness value is calculated for each of the possible pose sequences in each hypothesis of MHT. It is based on a simple line connecting two previous frames of the joint sequences. Such fitness values in each hypothesis of MHT are then ordered. The configuration corresponding to the best fitness is selected as the solution for that frame, while k best configurations according to their fitness values are kept for generating new hypotheses of MHT for the next frame. All trajectories of the left wrist in the walking, ball-throwing and back-flipping scenes are shown in Figures 6.18 (a), 6.19 (a) and 6.20 (a), respectively. It can be seen that multiple trajectories were obtained. This complicates the selection of the optimal solution. The trajectories of the left wrist after applying the joint-angle-limit constraint are shown in Figures 6.18 (b), 6.19 (b) and 6.20 (b). The trajectory of the ground truth and the optimal solution under the smoothness constraint are plotted by solid line and square makers, respectively.

It can be seen that our approach shows accurate results for both simple and strong perspective-effect cases. Moreover, the ground truth is always found (with very little uncertainty) within the set of configurations computed from our reconstruction of the 3D human pose. MHT can identify the optimal solution for all images in each image sequence as shown by overlapped parts between solid lines and square makers in Figures 6.18 (b), 6.19 (b) and 6.20 (b).

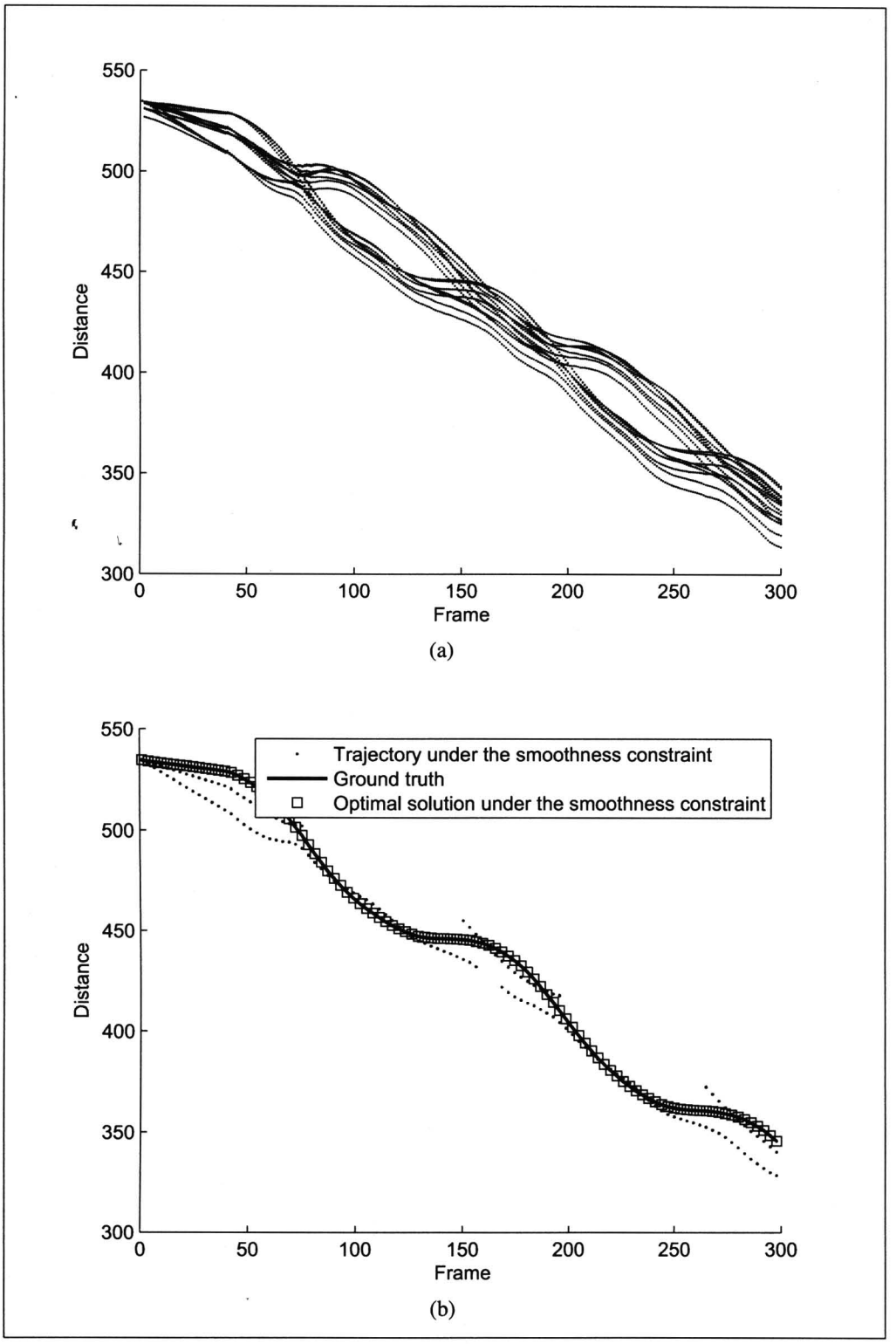

Figure 6.18 Trajectories of the left wrist in the walking (a) without employing a joint-angle limitation constraint and (b) after tracking

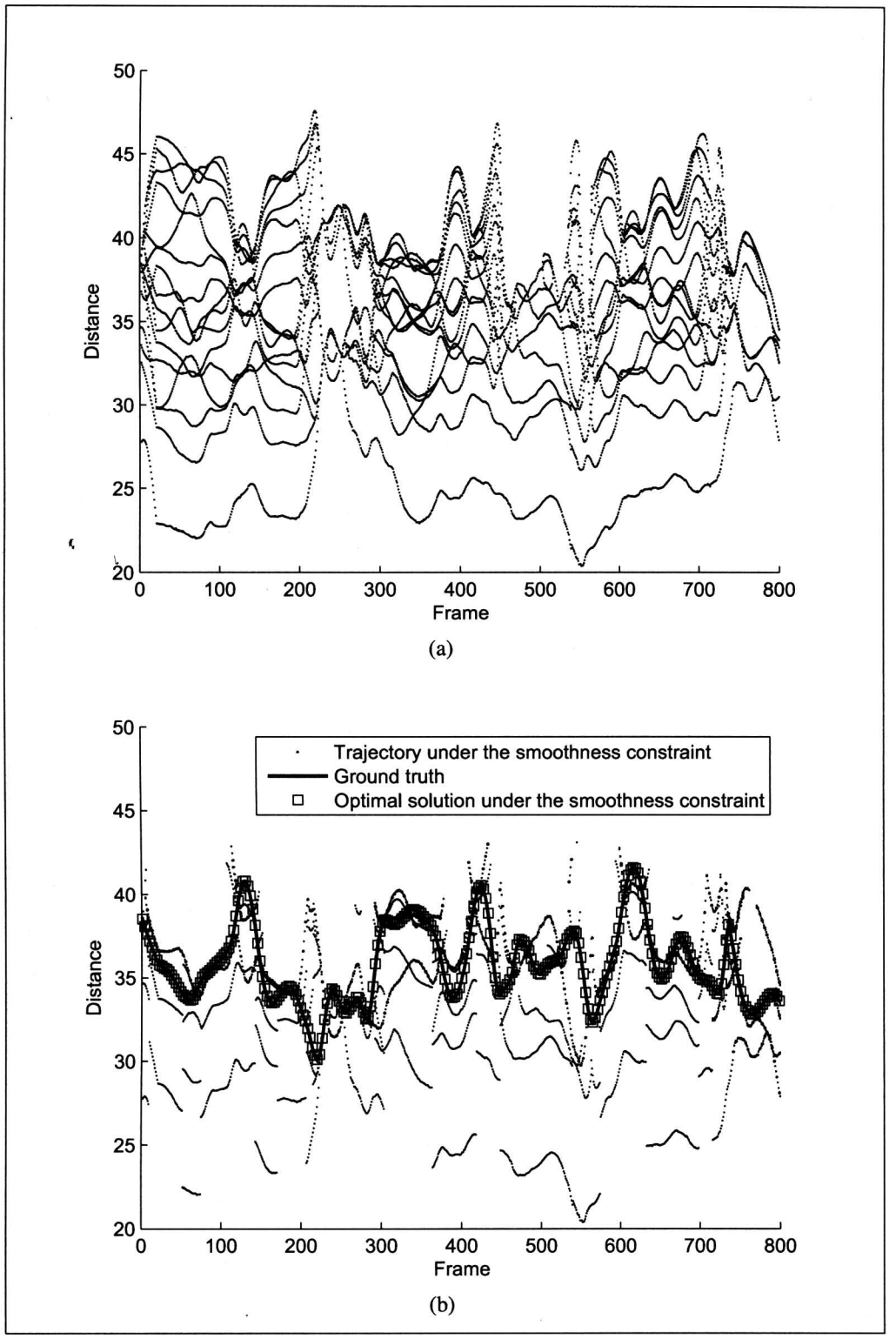

Figure 6.19 Trajectories of the left wrist in the ball-throwing (a) without employing a joint-angle limitation constraint (b) after tracking

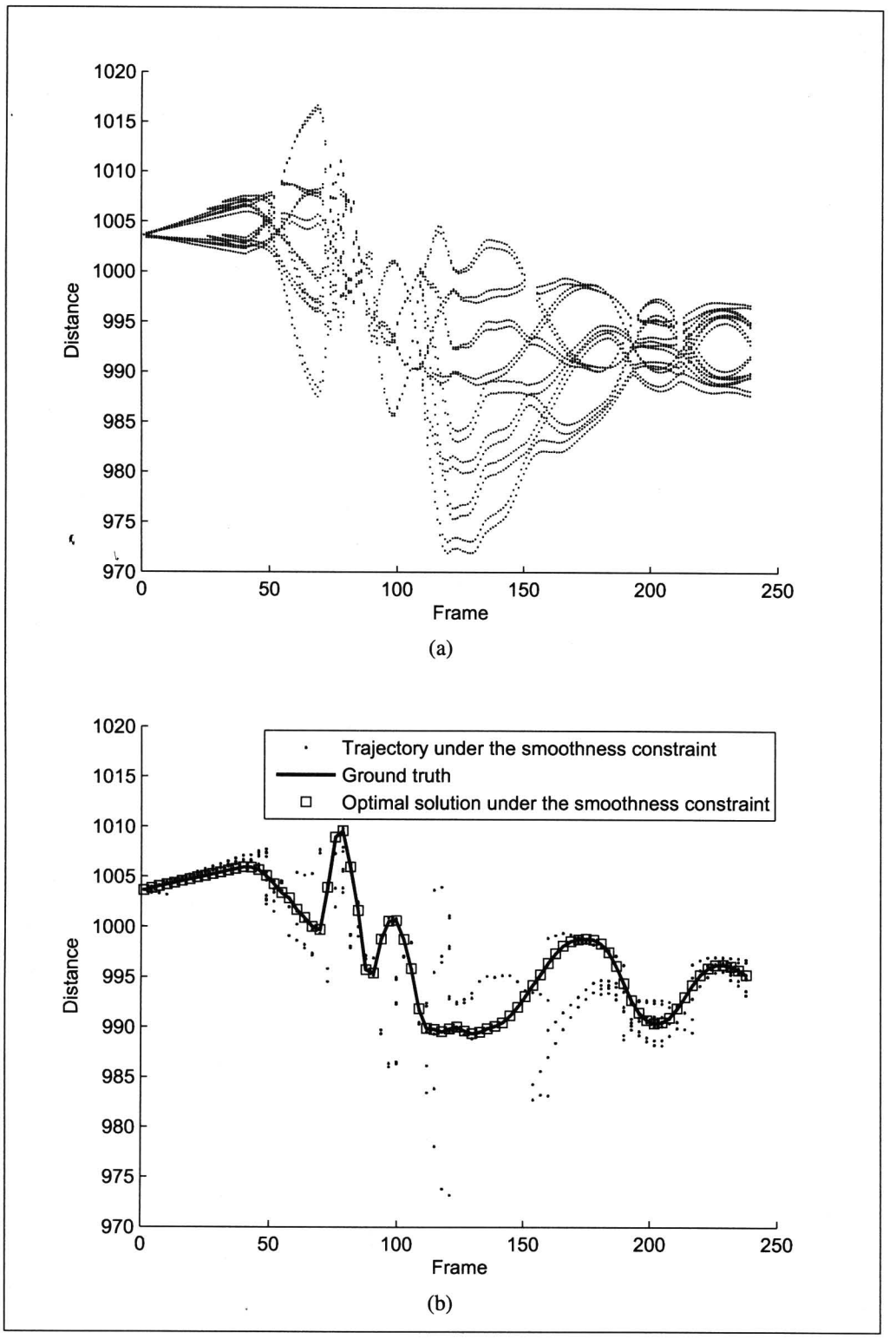

Figure 6.20 Trajectories of the left wrist in the back-flipping sequences (a) without employing a joint-angle limitation constraint (b) after tracking

To evaluate the performance of the smoothness function, we compared our smoothness function with smoothness function on the nearest approach on 1000 best hypotheses of MHT. The error distances of 3D human pose results compared with 3D real world coordinates of the ground-truth is shown in Table 6.4. It can be seen that our smoothness function is adequate and very efficient that the nearest based approach.

Table 6.4 The error distances (in centimeters) of our smoothness function and smoothness approach based on the nearest point approach

Scene	Our smoothness function	Smoothness function on the nearest point approach
Walking scene	1.25	2.11
Ball throwing scene	1.21	5.7
Back flip scene	1.97	3.2

6.4.4 Robustness against Errors in Reference Distance Determination

In this section, we investigated robustness of our approach due to errors in the reference distance estimation. The reference distance is computed in the first frame and used in the other frames. In our experiment, the reference distance R is varied between 70% to 130%. The average distance errors from the corresponding ground truth are shown in Figure 6.21. The average distance errors are between 1 to 6 centimeters per joint. By comparing the error to the human height of approximately 175 centimeters, the method is rather robust against errors in the reference distance determination.

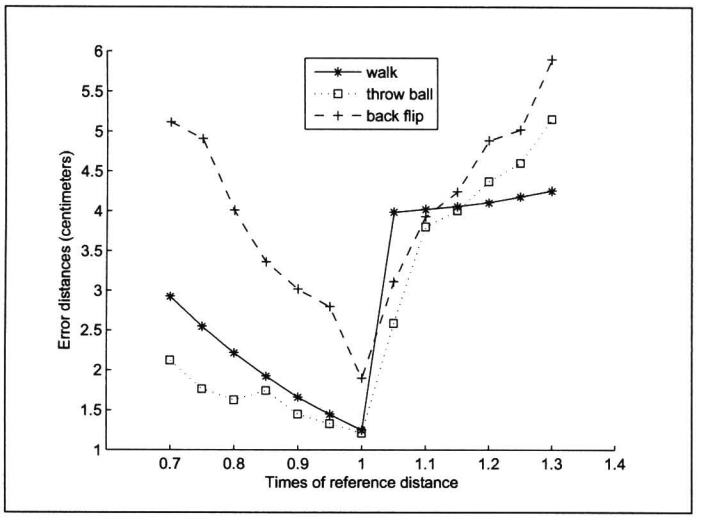

Figure 6.21 Error distance due to errors in difference reference distance determination

6.4.5 Robustness against Noise in 2D Data

Additionally, we investigated robustness of our proposed method on noise in the 2D data. We used synthesized data of 2D joint locations corrupted by a Gaussian noise. 2D Gaussian distributions with zero mean and variances 0.5, 1, 1.5, 2, 2.5 and 3 were generated and used in the experiments. The average distance errors from the corresponding ground truth are shown in Figure 6.22. It can be seen that our approach is robust against variation in the 2D input data.

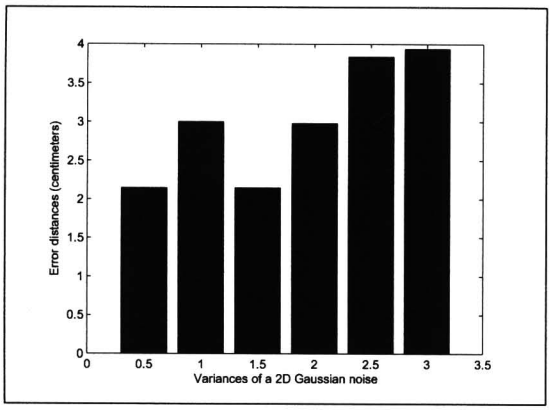

Figure 6.22 Average error distance due to errors in 2D joint locations

6.5 Conclusions

A novel method is presented to estimate 3D relative positions of an articulated body from 2D point correspondences in a single uncalibrated image. It is based on the perspective concept. It does not require any camera parameters from the user. Instead, our method computes a reference distance under a simpler assumption in the first frame. However, multiple configurations are obtained from the method because the problem is inherently non-uniqueness. An MHT-based tracking technique is introduced in order to select the best solution. This method is not only effective but also very efficient. Our approach was tested on both synthesized and real-world image sequences, and we obtained excellent results. Quantitative comparisons with scaled-orthographic [18] and previous perspective [25] approaches with respect to accuracy were also performed using image sequences with different degrees of perspective and pose complexities. Our proposed method outperformed these other approaches [18, 25], especially in scenes with strong perspective effects and various poses that are generally difficult to model.