Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Automatic reconstruction of 3D human motion pose from uncalibrated monocular video sequences based on markerless human motion tracking

Beiji Zou, Shu Chen*, Cao Shi, Umugwaneza Marie Providence

School of Information Science and Engineering, Central South University, Changsha 410083, People's Republic of China

ARTICLE INFO

Article history:
Received 19 December 2007
Received in revised form 3 September 2008
Accepted 17 December 2008

Keywords:
3D human motion reconstruction
Human motion tracking from monocular video sequences
Inverse kinematics
Human skeleton model
Camera model

ABSTRACT

We present a method to reconstruct human motion pose from uncalibrated monocular video sequences based on the morphing appearance model matching. The human pose estimation is made by integrated human joint tracking with pose reconstruction in depth-first order. Firstly, the Euler angles of joint are estimated by inverse kinematics based on human skeleton constrain. Then, the coordinates of pixels in the body segments in the scene are determined by forward kinematics, by projecting these pixels in the scene onto the image plane under the assumption of perspective projection to obtain the region of morphing appearance model in the image. Finally, the human motion pose can be reconstructed by histogram matching. The experimental results show that this method can obtain favorable reconstruction results on a number of complex human motion sequences.

1. Introduction

Human motion contains a wealth of information about actions, intentions, emotions, and personality traits of a person and plays an important role in many application areas, such as surveillance, human motion analysis, and virtual reality. It is a hot topic to track human joint and reconstruct the corresponding 3D human motion posture from an uncalibrated monocular video sequences, the human motion pose reconstruction can be categorized into two groups: (1) using multi-view video sequences, and (2) using monocular video sequences. Reconstruction of human motion pose from monocular video sequences is more attractive because it has many advantages such as convenient to use, conveniently available to general public and less restrictions. The depth value of an object will be lost when the object is projected onto 2D image plane. Therefore, 3D motion reconstruction from 2D motion sequences is still a challenging task. The conventional methods to reconstruct human pose from monocular video sequences may require some restrictions or prior knowledge. Rather than the classical algorithms, in this paper, we propose an approach to reconstruct the 3D human motion pose from uncalibrated monocular video sequences by combining human joint tracking and pose extraction, whose advantages include fewer constraints, without knowing the parameters of camera model, easy to implement and more precise performance of the pose reconstruction.

Human pose reconstruction from monocular video sequences is roughly divided into two categories: machine learning methods and object tracking methods. Researchers propose the use of machine learning methods that exploit prior knowledge in gaining more stable estimates of 3D human body pose [1–5]. However, these methods require a large amount of samples which limit their applications. Object tracking methods commonly follow two sequence steps: first, locating feature of human and tracking them in each frame, then, reconstructing human pose by these obtained features. Many researchers have conducted studies on the first step, and general surveys can be found in recent review papers [6,7], in this step, people always use the configuration in the current frame and a dynamic model to predict the next configuration [8,9]. Most approaches perform prediction by variants of kalman filtering [9,10] and particle filtering [11–13]. Particle filters restrict themselves to predictions returned by a motion model which is hard to construct, such a scheme is susceptible to drift due to imprecise motion model that the predictions were worse. Annealing the particle filter [14] or performing local searches [15] is the ways to attack this difficulty. The second step is human pose reconstruction (i.e., extracting 3D coordinates of feature from its corresponding 2D image coordinates). Some researchers reconstruct the human motion pose from video sequences by using some constrains such as human skeleton proportions based on camera model, these methods can be classified into two classes depending on the camera model adopted: (1) using affine camera model; and (2) using perspective camera model. Affine camera model is only an approximation of the real camera model. Scaled-orthographic camera model is an important instance of this kind and is popularly used

* Corresponding author.
E-mail address: csu_cs@163.com (S. Chen).

0031-3203/$ - see front matter Crown Copyright © 2009 Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.12.024

by many researchers [16,17]. The scale factor *s* has a significant effect on the result of human motion pose reconstruction by using scaled-orthographic camera model [16]. In these methods, the scale factor *s* is estimated by satisfying a constrained formula, but not a ground truth value; so the reconstructed human pose is great different from the real human pose, and these methods can only handle images with very little perspective effects. In addition, there are very limited research efforts working on human pose reconstruction based on perspective camera model [18–20]. Zhao et al. [20] restrict all body segments of the human figure as almost parallel to the image plane in order to acquire accurate human skeleton proportions. Peng requires estimating the virtual scale factor for each frame [19].

The remainder of this paper describes our algorithm in more detail. In Section 2, we explain the diagram of data flow in our system, and in Section 3, we describe the initialization of our system. The detail procedure to reconstruct human motion pose is described in Section 4, while in Section 5, we illustrate results from our system. In the end of this paper, we conclude about this study and point out the future further work.

2. Overview

The basic idea of our algorithm is to reconstruct 3D human pose reconstruction from the corresponding 2D joints on the image plane. The positions of human joints in each frame of the video are located with a local search by using the technique of morphing appearance model matching.

The proposed algorithm is divided into four major steps as shown in Fig. 1. The first step is to initialize models by a simple user interface with the first frame as input, the texture information and space information about the appearance of body segments can be acquired by marking projected landmarks of the subject's body on the image plane, and the relative lengths of body segments in the human skeleton model are also estimated.

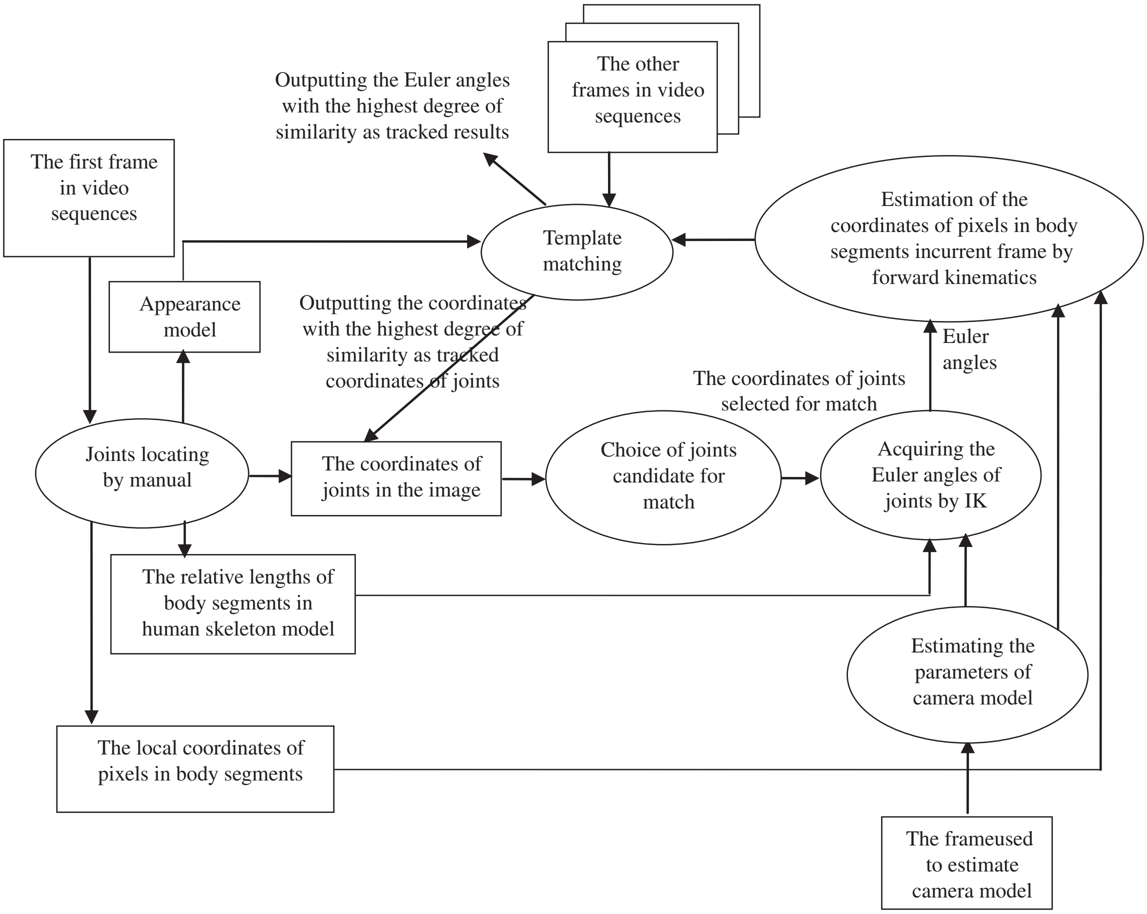
The other frames in video sequences
Outputting the Euler angles with the highest degree of similarity as tracked results
The first frame in video sequences
Template matching
Estimation of the coordinates of pixels in body segments incurrent frame by forward kinematics
Appearance model
Outputting the coordinates with the highest degree of similarity as tracked coordinates of joints
Euler angles
The coordinates of joints selected for match
Joints locating by manual
The coordinates of joints in the image
Choice of joints candidate for match
Acquiring the Euler angles of joints by IK
The relative lengths of body segments in human skeleton model
Estimating the parameters of camera model
The local coordinates of pixels in body segments
The frameused to estimate camera model

Fig. 1. Diagram of data flow in our system.

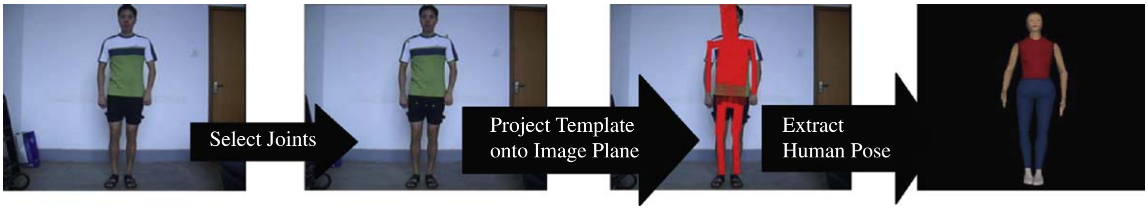

Fig. 2. Morphing appearance model matching to extract human pose.

Please cite this article as: B. Zou, et al., Automatic reconstruction of 3D human motion pose from uncalibrated monocular video sequences based on markerless human motion tracking. Pattern Recognition (2009), doi: 10.1016/j.patcog.2008.12.024

Next, the rotation Euler angles of its father joint are calculated by using inverse kinematics (IK) following two steps: (1) estimating the 3D coordinates of this joint by established camera model on the base of human skeleton proportions constrains, (2) calculating the Euler angles of its father joint by using IK based on the obtained 3D coordinates of this joint.

In the third step, the 3D coordinates of pixels in body segments of this joint are attached and after rotation they are calculated by forward kinematics. The rotated degree of segments are determined according to the Euler angles estimated in the second step, the projection of these pixels on the image plane consists of the morphing appearance model of body segments.

In the last step, the histogram matching is used to match the histogram of pixels in current frame with the histogram of pixels in the appearance model. Then, the joints with the highest degree of similarity are tracked as objects, and the human motion pose are reconstructed simultaneously (Fig. 2).

3. Building models

3.1. Human skeleton model

We represent human body as a tree stick model, which is inspired by the human body model employed at the Human Modeling and Simulation Center at University of Pennsylvania [21]. As shown in Fig. 3, the human skeleton model consists of rigid parts connected by joints, in which, J_1 is the root joint correspond to pelvis. Information about other joints is provided in Table 1. Fig. 4 shows the tree structure of human skeleton model. The relative lengths of human body segments in the model are ratios of lengths which can obtained from anthropomorphic measurement.

A local coordinate system is attached to each body part. The orientation of local coordinate system is shown in Fig. 3, and the origin of coordinates is located at the position of each joint. The kinematics is represented by a transformation tree whose root is the basic coordinate system and whose leaves are the coordinate systems of head, hands, and feet. The origin of the basic coordinate system is at the position of joint J_1, and the orientation of it is always the same as the local coordinate system of joint J_1 in initial state.

3.2. Appearance model of people

As for human motion tracking by template matching, the line is not a good selection for it contains little texture information. We approximate the limbs as planar regions for the reason that there is much color information on the body part, the appearance model of people consists of shape model and texture model as predicted in Fig. 5. According to the human skeleton model, the appearance model of people represented by 15 rectangles which correspond to body segments, each rectangle contains not only information of pixels inside it but also the coordinates of each pixel in local coordinate system. As shown in Fig. 5, the middle line of each rectangle is the skeleton in the human skeleton model, the middle point of the edge in each rectangle is the joint in the human skeleton model. In Fig. 5, we only marked joint J_1, the other joints can be marked similarly.

3.3. Camera model

3.3.1. Projection model

Under perspective camera model, the coordinates of a point in the scene, (x, y, z), can be related to the coordinates of its projection in the image, (u, v), through

$$\begin{pmatrix}u\\v\end{pmatrix} = \frac{1}{s}\begin{pmatrix}1&0&0\\0&1&0\end{pmatrix}\begin{pmatrix}x\\y\\z\end{pmatrix}. \tag{1}$$

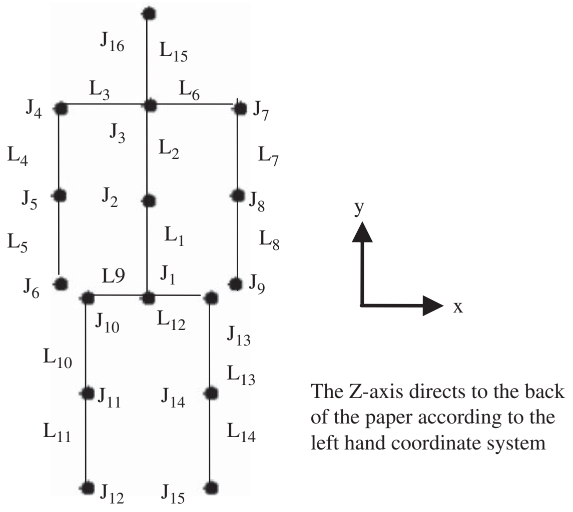
J16
L15
L3
L6
J4
J7
J3
L2
L4
L7
J5
J2
J8
L1
L5
L8
J6
L9
J1
J9
J10
L12
J13
L10
L13
J11
J14
L11
L14
J12
J15
y
x
The Z-axis directs to the back of the paper according to the left hand coordinate system

Fig. 3. The human skeleton model.

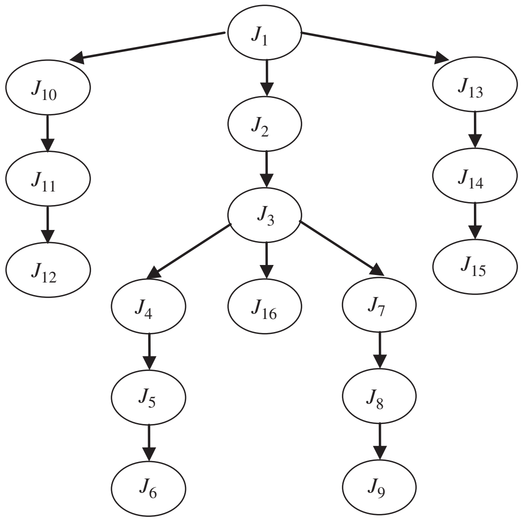

Fig. 4. The tree structure of human skeleton model.

Table 1
Information related to the joints of skeleton model.

ID	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8
Joint	Pelvis	Chest	Clavicle	Right shoulder	Right elbow	Right Wrist	Left shoulder	Left elbow
ID	J_9	J_{10}	J_{11}	J_{12}	J_{13}	J_{14}	J_{15}	J_{16}
Joint	Left wrist	Right hip	Right knee	Right ankle	Left hip	Left knee	Left ankle	head

Please cite this article as: B. Zou, et al., Automatic reconstruction of 3D human motion pose from uncalibrated monocular video sequences based on markerless human motion tracking. Pattern Recognition (2009), doi: 10.1016/j.patcog.2008.12.024

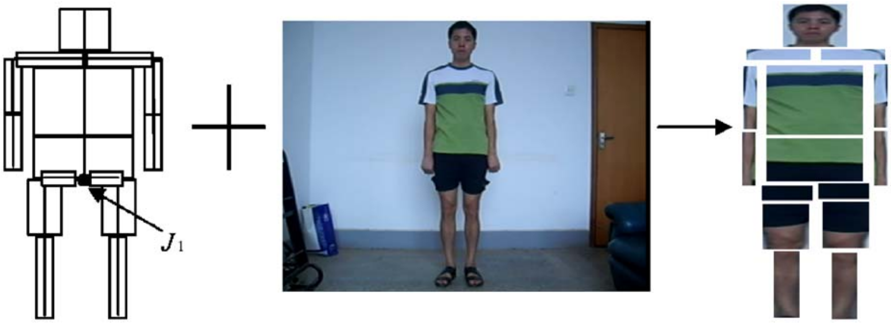

Fig. 5. This figure indicates how the appearance model of people built.

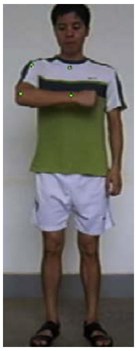

Fig. 6. A frame used to estimate the change of s corresponds to a unit change of z.

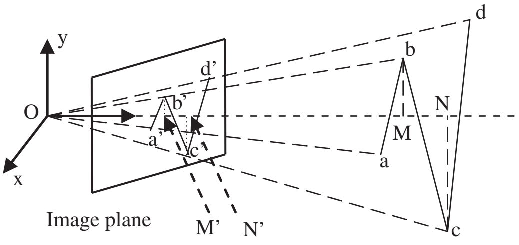

Fig. 7. The projection of three linked segments onto an image under perspective projection.

In this equation, the parameter s denotes an unknown scale factor which is defined as: $s=z/f$, where z is the z-coordinate of the point in the scene, f is the focal length of camera lens. From the definition of s, we know that when z-coordinate changes, the scale factor s will change linearly. The change of s, (ds), corresponding to the change of z, (dz), can be formulated as $T(dz)=ds$.

3.3.2. *Establishing the change of s corresponds to a unit change of z*

Firstly, we need to locate a frame from video sequences, in which three linked body segments with the first one and the last one parallel to the image plane. A general frame is shown in Fig. 6, in which the right shoulder and the right upper arm are parallel to image. Fig. 7 shows the projection of these three linked body segments, onto the image under perspective projection.

In this case, the segment ab and the segment cd are parallel to the image plane, the projection of these two segments onto the image are represented by $a'b'$ and $c'd'$, respectively. The segment bc is not parallel to the image plane, the corresponding projection onto image is represented by $b'c'$. The lengths of ab, bc and cd are denoted as L_{ab}, L_{bc} and L_{cd}, respectively, which can be obtained from the performance's body segment proportions. The lengths of $a'b'$, $b'c'$ and $c'd'$ are denoted as $L_{a'b'}$, $L_{b'c'}$ and $L_{c'd'}$, respectively, and these values can be calculated as follows, take $a'b'$ for example.

$L_{a'b'}=\sqrt{(a'_x-b'_x)^2+(a'_y-b'_y)^2}$, where (a'_x,a'_y), (b'_x,b'_y) are the coordinates of points a' and b' in the image, respectively.

The coordinates of point a in the scene are (a_x,a_y,a_z), and the coordinates of other points in Fig. 7 are denoted similarly.

As the segment ab is parallel to the image plane, the projection of segment ab onto Z-axis focus to the point M, the scale factor s corresponds to the point M can be calculated as $s_{ab}=OM/f=L_{ab}/L_{a'b'}$, similarly, we get the corresponding scale factor s to the projection point N of segment cd onto Z-axis as $s_{cd}=L_{cd}/L_{c'd'}$. As the segment ab and the segment cd are both parallel to the image plane,the distance between the point M and the point N is dz which subject to the following equation:

$$dz=c_z-b_z. \tag{2}$$

According to the space geometry knowledge, L_{bc} satisfies the following equation:

$$\begin{aligned}&\sqrt{(c_x-b_x)^2+(c_y-b_y)^2+(c_z-b_z)^2}=L_{bc}\\&\quad\Rightarrow\sqrt{(c_x-b_x)^2+(c_y-b_y)^2+\mathrm{d}z^2}=L_{bc}\\&\quad\Rightarrow\sqrt{(s_{cd}c'_x-s_{ab}b'_x)^2+(s_{cd}c'_y-s_{ab}b'_y)^2+\mathrm{d}z^2}=L_{bc}.\end{aligned} \tag{3}$$

As s_{cd}, s_{ab}, c'_x, b'_x, c'_y, b'_y and L_{bc} are known; dz can be calculated by Eq. (3). The absolute change of s, $(|ds|)$, corresponding to the absolute change of z, $(|dz|)$, can be calculated as follows:

$$|ds|=abs(s_{cd}-s_{ab}). \tag{4}$$

3.3.3. *The relative 3D coordinates of segments estimation*

Given the scale factor s of joint J_i is known as s_{J_i}. In the subsection we will explain how to estimate the 3D coordinates of joint J_{i+1} in camera projection space.

According to the invariable length of segment, we can establish the following equation:

$$\begin{aligned}&\sqrt{(J_i^x-J_{i+1}^x)^2+(J_i^y-J_{i+1}^y)^2+(J_i^z-J_{i+1}^z)^2}=L_i\\&\quad\Rightarrow\sqrt{(J_i^x-J_{i+1}^x)^2+(J_i^y-J_{i+1}^y)^2+(\Delta z)^2}=L_i\\&\quad\Rightarrow\sqrt{[s_{J_i}\cdot J_{i'}^x-(s_{J_i}+\Delta s)\cdot J_{i+1'}^x]^2+[s_{J_i}\cdot J_{i'}^y-(s_{J_i}+\Delta s)\cdot J_{i+1'}^y]^2+(\Delta z)^2}=L_i\\&\quad\Rightarrow\sqrt{[s_{J_i}\cdot J_{i'}^x-(s_{J_i}+T(\Delta z))\cdot J_{i+1'}^x]^2+[s_{J_i}\cdot J_{i'}^y-(s_{J_i}+T(\Delta z))\cdot J_{i+1'}^y]^2+(\Delta z)^2}=L_i.\end{aligned} \tag{5}$$

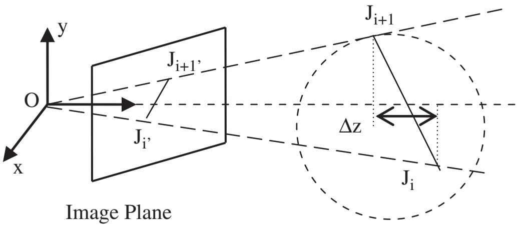

Fig. 8. The projection of an articulated object onto an image under perspective projection.

where L_i is the relative length of the segment L_i. (J_i^x, J_i^y, J_i^z), $(J_{i+1}^x, J_{i+1}^y, J_{i+1}^z)$ are the coordinates of J_i and J_{i+1} in the projection space, respectively. $(J_{i'}^x, J_{i'}^y)$, $(J_{i+1'}^x, J_{i+1'}^y)$ are the coordinates of their projection of J_i and J_{i+1} in the image, respectively. Δz is the relative depth between J_{i+1} and J_i, $\Delta z = J_{i+1}^z - J_i^z$. Eq. (5) is a quadratic equation of one variable Δz. Thus, there are three different kinds of solutions for this equation: only one real number, two real numbers, and imaginary number, according to the relative position between line and sphere in projection space. In case of two real numbers, we should choose one of these two real numbers as the ground truth. We establish the projection coordinate system as in Fig. 8, the origin of coordinate axes is O, Z-axis through the center of image and the positive direction direct to object. According to definitions of Δz and s, we choose Δz as a negative number while J_{i+1} is nearer than J_i to the image plane. Otherwise, we choose Δz as a positive number. Fig. 8 shows the situation of intersection of the line and the sphere, in this case, Δz is only one negative real number.

Since we have calculated Δz from Eq. (5), the scale factor $s_{J_{i+1}}$ corresponds to the joint J_{i+1} can be obtained as $s_{J_{i+1}} = s_{J_i} + T(\Delta z)$.

The coordinates of J_i and J_{i+1} in the projection space can be calculated as

$$\begin{aligned} J_i^x &= s_{J_i} \cdot J_{i'}^x, \\ J_i^y &= s_{J_i} \cdot J_{i'}^y, \\ J_{i+1}^x &= s_{J_{i+1}} \cdot J_{i+1'}^x, \\ J_{i+1}^y &= s_{J_{i+1}} \cdot J_{i+1'}^y. \end{aligned} \tag{6}$$

4. Initialization

The problem can be decomposed into three sub-problems: the estimation of the relative lengths of body segments in the human skeleton model, the initialization of appearance model of people and the estimation of the scale factor s correspond to root joint.

We have developed a graphical user interface that allows the user to select the projection of joints of the subject's body in the first frame. A marked image is shown in Fig. 9, in which the green dots depict all selected joints while the yellow rectangles depict the position of body segments. The body configuration in the first frame is assumed as follows: human in a standing position, facing the $+Z$ direction, the arms should be straight and parallel to the sides of the body. As shown in Fig. 9, the color information of pixels in these rectangles is served as the color model for the matching of subsequent frames used in Section 5.3.

4.1. Estimation of the relative lengths of body segments in the human skeleton model

Although the relative lengths of skeletons obtained by anthropometry are reliable, the positions of marked joints in the image may not be the ground truth. So we estimate the proportions of performer's skeletons by refer to the length of segment L_5.

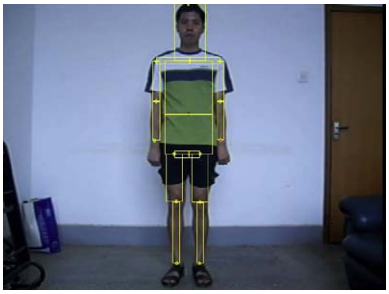

Fig. 9. The image to indicate the marked result.

We set L_5 as the length of segment L_5 which can be obtained by anthropometric measurements. The length of L_5 projected onto image plane is denoted as $L_{5'}$ which can be calculated as $L_{5'} = \sqrt{(J_{5'}^x - J_{6'}^x)^2 + (J_{5'}^y - J_{6'}^y)^2}$, where $(J_{5'}^x, J_{5'}^y)$ and $(J_{6'}^x, J_{6'}^y)$ are the image coordinates of J_5 and J_6. As the whole body parallels to the image plane, all joints in the image correspond to a same scale factor s: $s = L_5/L_{5'}$.

As the lengths of other segments in the image $L_{i'}$ is calculated, the relative lengths of these segments in the human skeleton model can be obtained as $L_i = s \cdot L_{i'}$.

4.2. Initialization of the appearance model of people

As we explained in Section 3.2 that the appearance of people consists of pixels in body segments, the value of pixels can obtained from marked rectangle in the image. The spatial information about pixels is represented by their local coordinates. Given the local image coordinates of a pixel are (x, y), the local coordinates of this pixel in the scene can be obtained as $(s \cdot x, s \cdot y, 0)$.

5. Human motion tracking by template matching

The human motion tracking is performed with a local search in the image by template matching. The whole tracking is decomposed into three steps: (1) the estimation of rotation Euler angles based on the coordinates of joint candidate, (2) estimating the coordinates of the pixels of appearance model in the scene after rotation by forward kinematics based on estimated Euler angle, and (3) estimating the region of morphing appearance model by projecting these pixels onto the image plane, then reconstructing the human motion pose by histogram matching.

The 3D human motion pose and projected location of joints on the image plane are estimated in depth-first order as shown in Fig. 4. The coordinates of the child joint should be estimated with local search after estimation of the coordinates of the father joint. So the key to estimate the human motion pose is to estimate the coordinates of root joint firstly.

As we know, an adjacent area with the root joint at the center will not deform during human motion, we can estimate the coordinates of root joint with local search by template matching, and the template is established by the rectangle with root joint at the center.

In our system, we assume that joint J_1 moves parallel to the image plane without Z direction displacement. So that the scale factor s calculated in Section 4.1 can be used as the scale factor s of root joint in whole video sequences.

5.1. Estimating rotation Euler angles

Given the image coordinates of a joint candidate are known, the scale factor s of its father joint is also estimated. We can estimate the depth difference Δz between these two joints and their coordinates of x, y in the scene as explained in Section 3.3.3.

As shown in Fig. 3, the local coordinates of joints can be obtained as follows: the vector of J_2 in J_1 local coordinate system is [0, L_1, 0], the vector of J_3 in J_2 local coordinate system is $[0, L_2, 0]$, the vector of J_4 in J_3 local coordinate system is $[-L_3, 0, 0]$. The vector of other joints can be obtained similarly.

We assume that the order of joint rotation as Z–Y–X Euler angles, α is around the Z-axis, β is around the Y-axis, γ is around the X-axis.

Step 1: Estimation of the Euler angles of joint J_1. The function of torsional moment is not considered in 3D data. As a result, the motion component around the Y-axis is zero and β=0. Let ${}^{0}_{1}R$ denotes the rotation matrix that transforms a vector in J_1 local coordinates into basal coordinates, which is defined as

$$\begin{aligned}{}^{0}_{1}R &= \begin{bmatrix}\cos\alpha & -\sin\alpha & 0\\ \sin\alpha & \cos\alpha & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}1 & 0 & 0\\ 0 & \cos\gamma & -\sin\gamma\\ 0 & \sin\gamma & \cos\gamma\end{bmatrix}\\ &= \begin{bmatrix}\cos\alpha & -\sin\alpha\cos\gamma & \sin\alpha\sin\gamma\\ \sin\alpha & \cos\alpha\cos\gamma & -\cos\alpha\sin\gamma\\ 0 & \sin\gamma & \cos\gamma\end{bmatrix}.\end{aligned} \tag{7}$$

According to the kinematics knowledge, the relation between local coordinates and projection coordinates of L_1 is given by the following equation:

$${}^{0}_{1}R\begin{bmatrix}0\\ L_1\\ 0\end{bmatrix} = \begin{bmatrix}J_2^x - J_1^x\\ J_2^y - J_1^y\\ J_2^z - J_1^z\end{bmatrix} = \begin{bmatrix}J_2^x - J_1^x\\ J_2^y - J_1^y\\ \Delta z\end{bmatrix}, \tag{8}$$

where the vector $[0, L_1, 0]$ is the coordinates of J_2 in J_1 local coordinate system, the vector $[J_2^x - J_1^x, J_2^y - J_1^y, J_2^z - J_1^z]$ is the coordinates of J_2 in the basic coordinate system.

Substituting ${}^{0}_{1}R$ into the above equation yields equations as:

$$\begin{cases}\sin\alpha\cos\gamma = \frac{J_1^x - J_2^x}{L_1},\\ \cos\alpha\cos\gamma = \frac{J_1^y - J_2^y}{L_1},\\ \sin\gamma = \frac{\Delta z}{L_1}.\end{cases} \tag{9}$$

By solving it, the value of α and γ can be obtained as

$$\gamma = \arcsin\left(\frac{\Delta z}{L_1}\right), \quad \alpha = -\arctan\left(\frac{J_2^x - J_1^x}{J_2^y - J_1^y}\right). \tag{10}$$

Step 2: The estimation of the Euler angles of joint J_i. We assume that the rotation matrix that transforms a vector in J_{i-1} local coordinates into basal coordinates has been obtained.

$${}^{0}_{i-1}R = {}^{0}_{1}R\,{}^{1}_{2}R\,{}^{2}_{3}R \ldots {}^{i-2}_{i-1}R, \tag{11}$$

where ${}^{0}_{1}R$, ${}^{1}_{2}R$, etc. are the rotation matrices between two coordinate system which attached to two consecutive joints in the access route from J_1 to J_{i-1}.

The relation of coordinates of J_{i+1} in J_i local coordinate system and coordinates of J_{i+1} in basic coordinate system can be written as

$$\begin{aligned}&{}^{0}_{i}RP_{i+1} = P\\ &\quad\Rightarrow {}^{0}_{i-1}R^{i-1}_{i}RP_{i+1} = P,\\ &P = \begin{bmatrix}J_{i+1}^x - J_i^x\\ J_{i+1}^y - J_i^y\\ J_{i+1}^z - J_i^z\end{bmatrix},\end{aligned} \tag{12}$$

where P_{i+1} is the coordinates of J_{i+1} in J_i local coordinate system which has been explained in the beginning of this section, P is the coordinates of J_{i+1} in basic coordinate system (only rotation, not include displacement), and ${}^{i-1}_{i}R$ is the rotation matrix that used to rotate the J_i local coordinate system to the J_{i-1} local coordinate system. ${}^{i-1}_{i}R$ can be represented in two forms. In the case of segment L_i perpendicular to X-axis, ${}^{i-1}_{i}R$ is defined as ${}^{0}_{1}R$, just like L_2, L_4, L_7, etc.; otherwise, due to the motion component around the X-axis is zero, thus, $\gamma = 0$, and ${}^{i-1}_{i}R$ is defined as

$${}^{i-1}_{i}R = \begin{bmatrix}\cos\alpha & -\sin\alpha & 0\\ \sin\alpha & \cos\alpha & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}\cos\beta & 0 & \sin\beta\\ 0 & 1 & 0\\ -\sin\beta & 0 & \cos\beta\end{bmatrix}\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}. \tag{13}$$

As L_3 is parallel to X-axis, ${}^{i-1}_{i}R$ used in the reconstruction of right clavicle pose is a case in point. Others include L_6, L_9, and L_{12}.

By solving Eq. (12), the Euler angles of J_i can be obtained.

5.2. Estimating the projected region of segments after rotation

We assume that the Euler angles calculated in Section 5.1 are (α, β, γ), the coordinates of its father joint in the scene are (J_{i-1}^x, J_{i-1}^y), and the scale factor s of its father joint is $s_{J_{i-1}}$. The coordinates of pixels in the corresponding body segment in the scene after rotation can be determined as follows:

$$\begin{aligned}\begin{bmatrix}x_r\\ y_r\\ z_r\end{bmatrix} &= \begin{bmatrix}\cos\alpha & -\sin\alpha & 0\\ \sin\alpha & \cos\alpha & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}\cos\beta & 0 & \sin\beta\\ 0 & 1 & 0\\ -\sin\beta & 0 & \cos\beta\end{bmatrix}\\ &\quad\times\begin{bmatrix}1 & 0 & 0\\ 0 & \cos\gamma & -\sin\gamma\\ 0 & \sin\gamma & \cos\gamma\end{bmatrix}\begin{bmatrix}x_l\\ y_l\\ 0\end{bmatrix}.\end{aligned} \tag{14}$$

Given the local coordinates of pixels in the appearance model are $(x_l, y_l, 0)$ which can be obtained from initialization as explained in Section 4.2, the local coordinates of pixels after rotation can be obtained by the above equation.

So the x, y coordinates of pixels in the scene can be calculated as

$$\begin{bmatrix}x_p\\ y_p\end{bmatrix} = \begin{bmatrix}J_{i-1}^x\\ J_{i-1}^y\end{bmatrix} + \begin{bmatrix}x_r\\ y_r\end{bmatrix}. \tag{15}$$

The scale factor s of pixels is calculated as $s_p = s_{J_{i-1}} + T(z_r)$.

Therefore, the image coordinates of these pixels can be obtained as $(x_p/s_p, y_p/s_p)$.

5.3. Template matching

As explained in Section 4, the color information of pixels in this segment can be obtained, we define the color histogram of it as a reference one which is denoted as H_1. From the above explanation, given a joint candidate, we can obtain the projected region of this body segment on the image plane based on the location of this joint candidate. Fig. 10 shows the result, and we define the color histogram of the projected region as a comparative one which is denoted as H_2. The similarity between the reference histogram and the comparative one is defined as

$$d(H_1, H_2) = \sqrt{1 - \sum_i \sqrt{H_1(i) \cdot H_2(i)}}, \tag{16}$$

where $H_1(i)$, $H_2(i)$ are the corresponding component in H_1 and H_2, respectively.

The joint candidate which satisfies the following formula will be determined as the expected one, and the estimated Euler angles corresponding to this joint candidate will be the reconstructed pose:

$$\widehat{p} = \arg\min\{d(H_1, H_i), i \in A\}, \tag{17}$$

Please cite this article as: B. Zou, et al., Automatic reconstruction of 3D human motion pose from uncalibrated monocular video sequences based on markerless human motion tracking. Pattern Recognition (2009), doi: 10.1016/j.patcog.2008.12.024

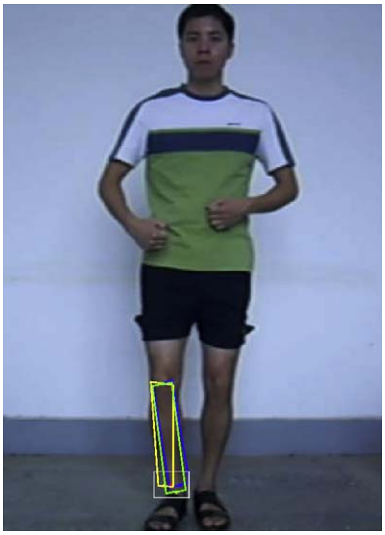

Fig. 10. This figure indicates the result of projected right leg block on the image plane under three joint candidates and determination of the expect one by local search. The dot rectangle indicates the search region, the red dot indicates the search center, and the green dot is the expected location of the joint. The corresponding Euler angles are estimated as the pose of this joint. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where A is the search region defined as

$$A = \{(x,y), |x - x_0| \leqslant h, |y - y_0| \leqslant h\}, \tag{18}$$

$h = 1.5 \times \max\{|x_i - x_{i-1}|, |y_i - y_{i-1}|\}$, where (x_i, y_i), (x_{i-1}, y_{i-1}) are the coordinates of this joint on the image plane estimated in the last frame and in the frame before last frame. (x_0, y_0) is the search center which can be obtained by the projection of this joint on the image plane as explained in Section 5.2. In this situation, the input Euler angles (α, β, γ) are set as the rotation Euler angles of this joint estimated in the last frame, and $(x_l, y_l, 0)$ are set as the local coordinates of this joint as explained in Section 5.1. The estimated image coordinates of this joint are obtained as (x_0, y_0).

Based on the estimated Euler angles, the x,y coordinates in the scene and scale factor s of this joint can be obtained by the method explained in Section 5.2. Then they are fed to the next procedure as inputs to estimate the rotation Euler angles of its son joint.

6. Experiments with real sequences

To test the proposed contribution, we measure 3D human motion pose on the same subject while the joints in the image were manually marked or located by semi-automatic tracking. In the first experiment, we present results to show how the imprecise estimation of $T(dz)$ affects the result of human motion pose reconstruction, and discuss which factors affect the precise estimation of $T(dz)$. The second experiment is performed to test the effectiveness of the proposed 3D joint points estimation from the projection of these points on the image plane. In this experiment, the location of joints was manually marked. The third experiment shows the tracking for a test sequences in which the subject's left hand moves in a decreasing spiral. In the last experiment, we present the comparative results as compare the proposed method to the scaled-orthographic projection method. In these experiments, the corresponding reconstructed human poses are all demonstrated by a virtual human.

The test video sequences were captured by a stationary Samsung digital camera (S600) with a temporal sampling rate of $\Delta t = \frac{1}{30}$ s and the resolution of 640 × 480. During the sequences the player moves parallel to image plane without z displacement. The relative lengths of segments in the human skeleton model were manually obtained from the measurements of the performer's body. Table 2 indicates the values that were used for the various segments. In these experiments, as shown in the first image in Fig. 11, we select the right clavicle, the right upper arm, and the right lower arm as three linked body segments to estimate the change of s corresponds to a unit change of z. Four paper markers have been stuck on the location of joints in order to locate the joints correctly.

Table 2
Values used for the relative lengths of the segments in human figure.

Segment	Relative length
Lower torso (L_1)	20
Upper torso (L_2)	34
Right/left clavicle (L_3/L_6)	21
Right/left upper arm (L_4/L_7)	21
Right/left lower arm (L_5/L_8)	24
Right/left hip (L_9/L_{12})	10
Right/left thigh (L_{10}/L_{13})	30
Right/left leg (L_{11}/L_{14})	36
Neck (L_{15})	20

6.1. *How the precision on the estimation of $T(dz)$ affect the result of human pose reconstruction*

It is important to note that imprecision on the estimation of $T(dz)$ may greatly affect the results of human pose reconstruction. In this section, some experiments for quantifying the influence of this estimation on the final results are presented. As shown in Fig. 11, we use three different precise joints locations to estimate the change of s corresponds to a unit change of z. The first image indicates the marked joint location with the highest precision. In the second image, the location of joints varies a little from the ground truth, and in the third image, the pixel error between marked joints and the corresponding ground truth is very great. Fig. 12 shows the results of human pose reconstruction under these three different estimation of $T(dz)$. We observe that the more precision on estimation of $T(dz)$, the higher degree of similarity between the original human pose and the reconstructed one. Given the pixel error between the marked joints position and the ground truth, the precision on the estimation of $T(dz)$ will depend on some factors including camera resolution and the distance between the subject and the camera, etc. The higher resolution the digital camera is, the more precise the estimation of $T(dz)$ will be, and the nearer the subject close to the digital camera, the more precise the estimation of $T(dz)$ will be.

6.2. *Reconstructing the human pose while the location of joints are manually marked*

To test the effectiveness of our proposed 3D human pose reconstruction from the corresponding 2D joints on the image plane, in this experiment, the performer in the test video sequences was outfitted with markers which enable us to locate the joints of the figure more accurate, and the position of joints on the image plane were manually labeled. Fig. 13 depicts the very encouraging reconstructed results. From the results, we can see that the reconstruction generated is highly satisfiable.

6.3. *Reconstructing the human pose by automatic tracking*

In this experiment, the human motion tracking is done on the input video sequences without preprocessing, and presume that the rotation Euler angles of hips to be zero to eliminate the negative influence of inaccurate estimation of root joint position on the image

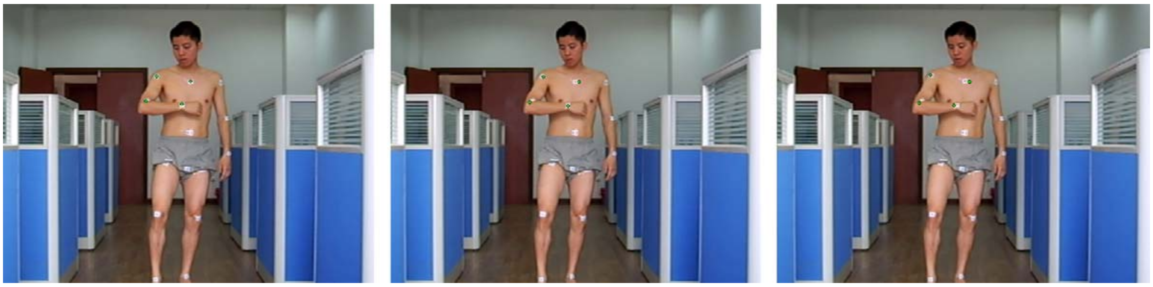

Fig. 11. Images to be used for estimation of the $T(dz)$. The corresponding change of s corresponds to a unit change of z in these three images were estimated as 0.0011730575008468, 0.0023854678824606 and 0.0049679396543650, respectively.

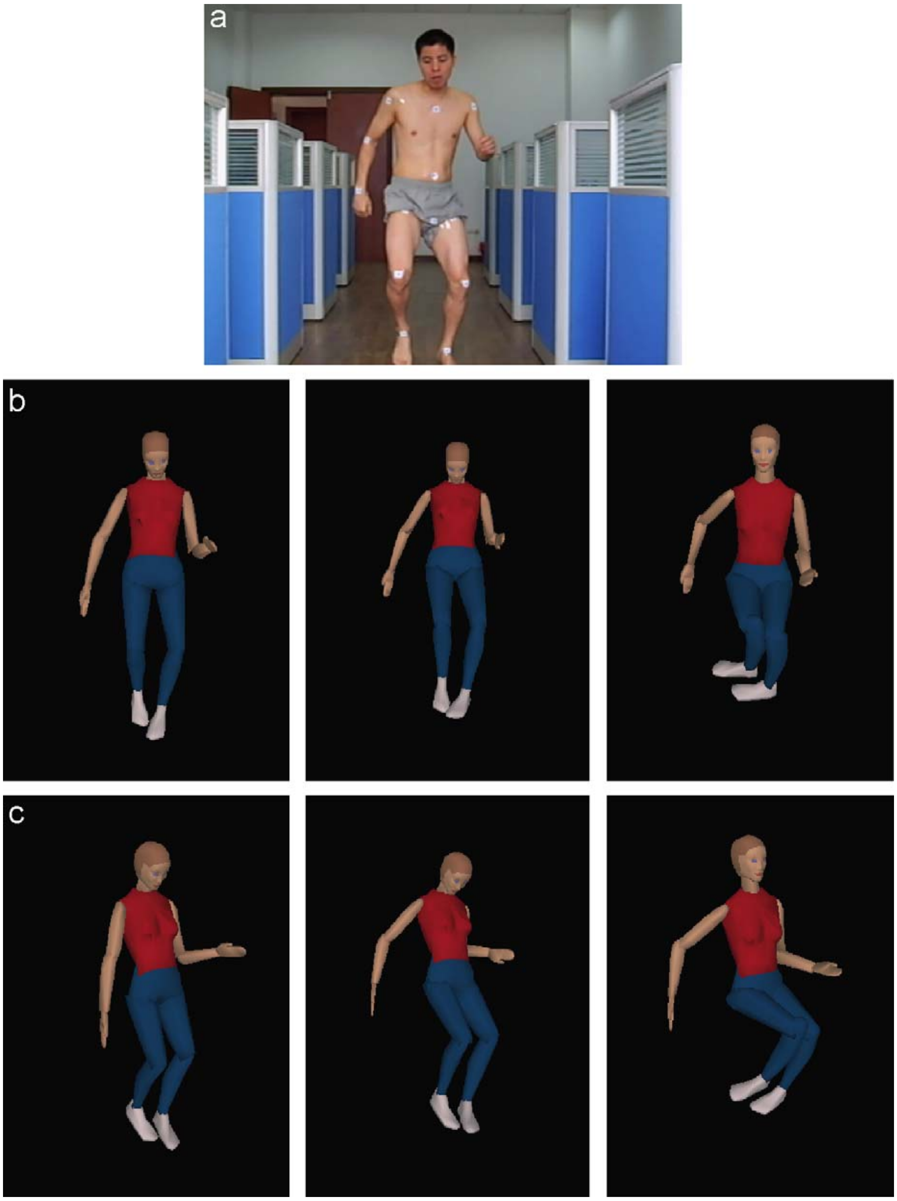

Fig. 12. Results of 3D human pose reconstructing under different estimation of $T(dz)$. The left most column is the result of reconstruction under the estimation of $T(dz)$ from the first image in Fig. 11, the middle column and the right most column are the results of reconstruction under the estimation of $T(dz)$ from the second image and the third image in Fig. 11, respectively. (a) Original image. (b) The reconstructed human pose from the front view. (c) The reconstructed human pose from the side view.

plane. The experimental results are based on a total of 300 video frames, and some of the reconstructed 3D human poses are presented in Fig. 14. Part (a) depicts the tracked frames, part (b) shows the rectangles of human body parts projected on the original images, part (c) shows reconstructed results simulated by the virtual human, and a rendered 3D side view is shown in the part (d) to illustrate

Please cite this article as: B. Zou, et al., Automatic reconstruction of 3D human motion pose from uncalibrated monocular video sequences based on markerless human motion tracking. Pattern Recognition (2009), doi: 10.1016/j.patcog.2008.12.024

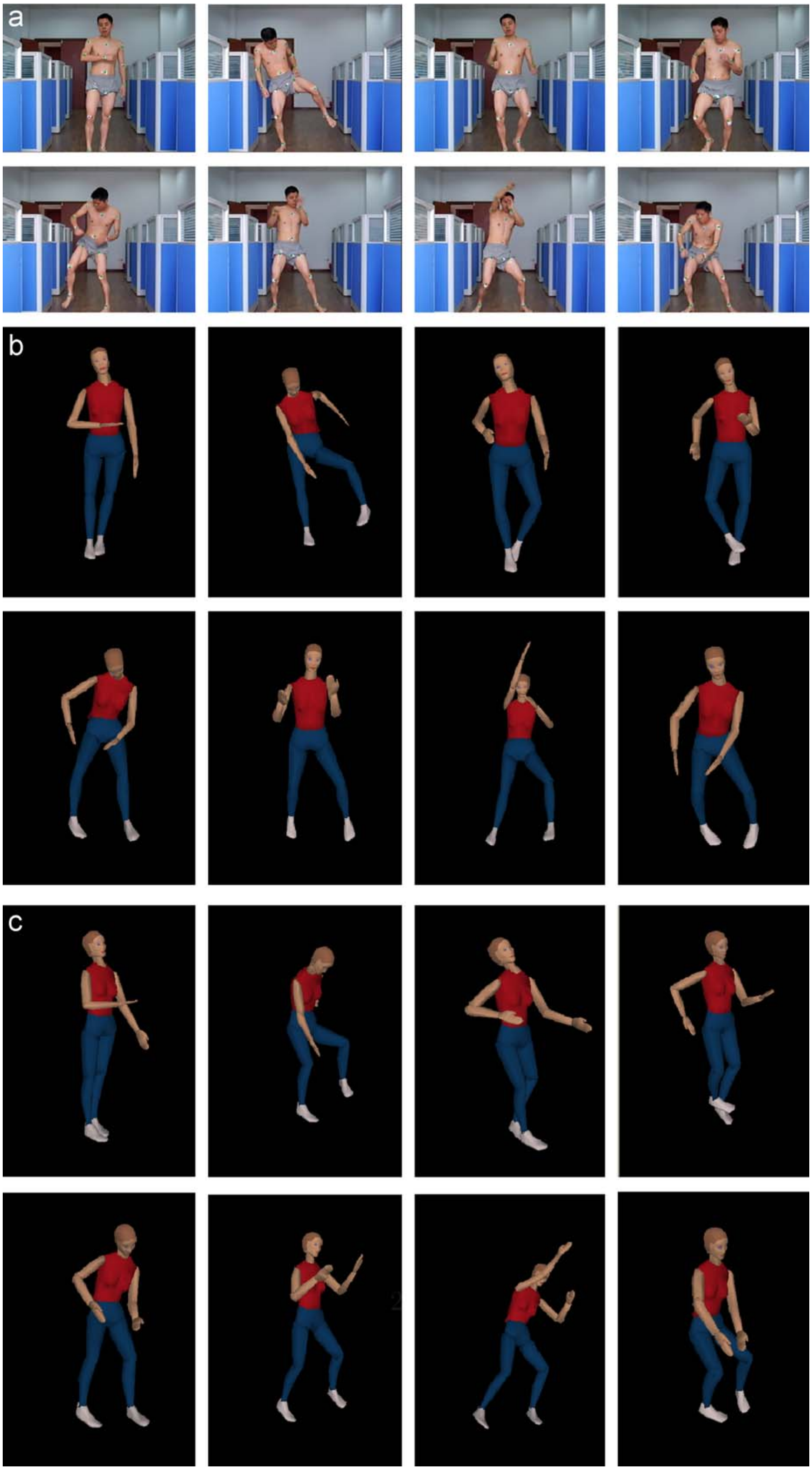

Fig. 13. 3D reconstruction of human pose by proposed algorithm. (a) Original images. (b) 3D reconstruction rendered from similar viewpoint. (c) Rendering from viewpoint rotated 45°.

the estimated relative depth. From the results, we can see that pose accuracy varies over frames, but rough body pose is for the most part visually accurate. The local errors in these reconstructed sequences, mostly due to illumination change and imprecision estimation of $T(dz)$. The error will increase when the segment is far from the root joint in human skeleton model, e.g., the estimated angle error of right

Please cite this article as: B. Zou, et al., Automatic reconstruction of 3D human motion pose from uncalibrated monocular video sequences based on markerless human motion tracking. Pattern Recognition (2009), doi: 10.1016/j.patcog.2008.12.024

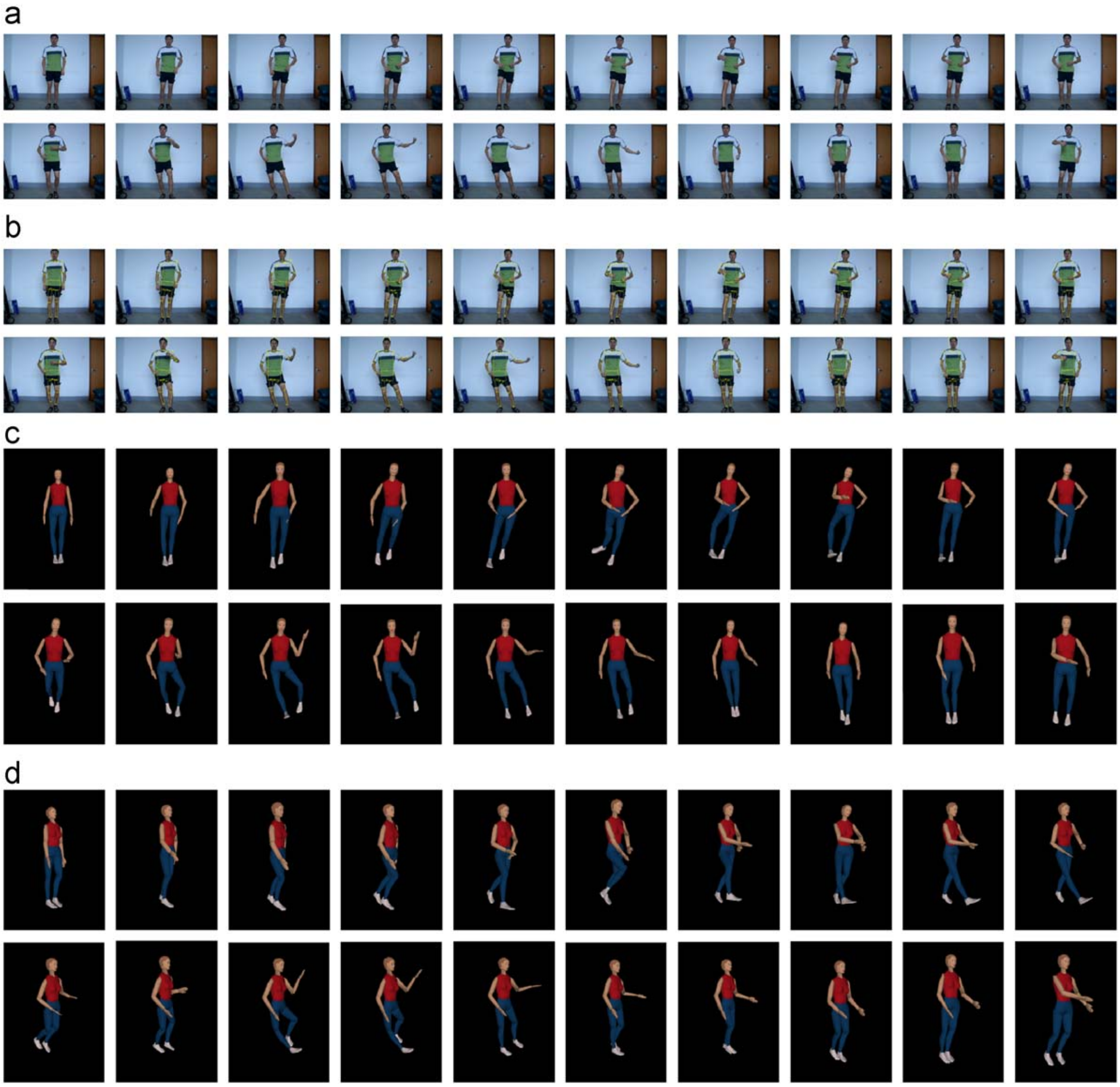

Fig. 14. Results of reconstruction of 3D human pose from the test video sequences in which the subject's left arm moves in a decreasing spiral (every 15th frame from 15 to 300 arranged from left to right, top to bottom). (a) Tracked frames. (b) The projection of expected model configuration on the original images. (c) Reconstructed human poses from the front view. (d) Reconstructed human poses from the side view.

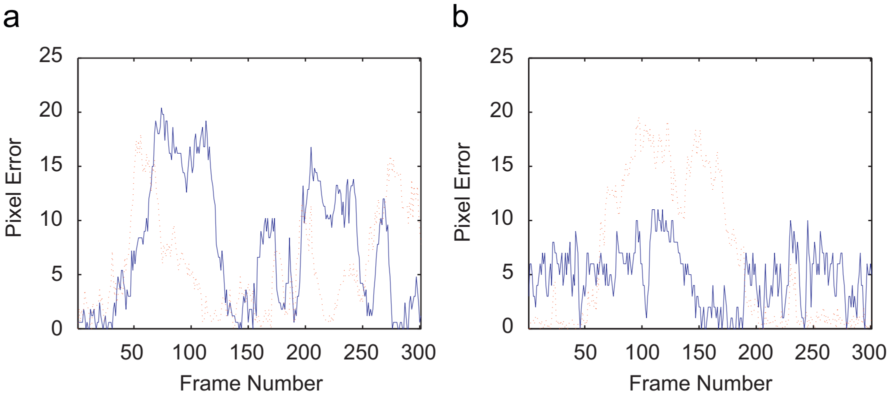

Fig. 15. Comparative plots of some joint locations obtained from tracking and manually labeled ground truth data. Blue solid plots indicate the pixel error in x-coordinate while red dot plots indicate the pixel error in y-coordinate. (a) J_9. (b) J_{12}. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

leg is bigger than the one of right thigh, and we notice that the error will be great when the human parts rotate drastically which can be seen on the estimated pose of right upper arm in frame 300th.

To quantify the accuracy of the tracking, we report the pixel error between the estimated location of joints in image and its corresponding, manually labeled, ground truth data in Fig. 15. Part (a)

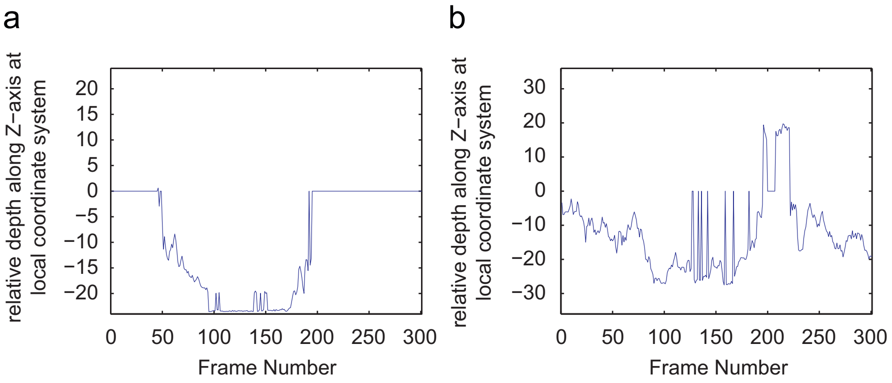

Fig. 16. The variation of relative depth value during human motion. (a) Left lower arm. (b) Right lower leg.

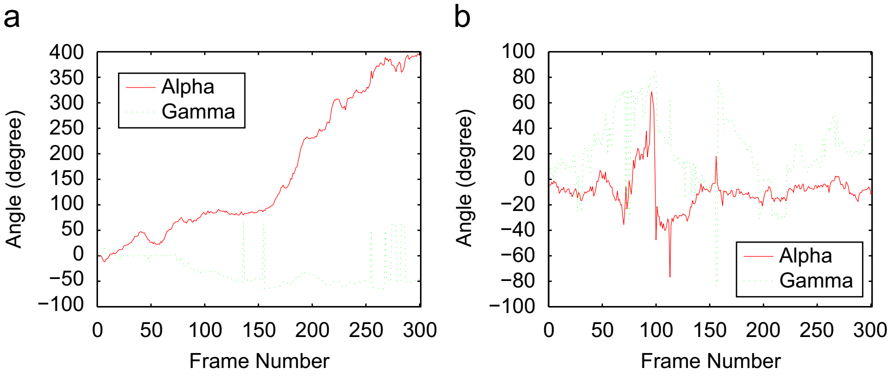

Fig. 17. The reconstructed motion trajectory of some human parts. Red solid plots indicate the estimated α angle of a segment while green dot plots indicate the estimated γ angle of the segment. (a) Left lower arm. (b) Right lower leg. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

shows the result of joint J_9 while part (b) shows the result of joint J_{12}. From these images, we can see that the most pixel error is limited to ±15. The peak of error is about from frame 50th to 150th; the reason is that the inaccurate estimated pose of lower torso and upper torso greatly affect the estimation of 2D position of these joints.

Fig. 16 depicts the whole variation of relative depth value (dz) during the human motion. The obtained dz is the relative value under human skeleton model as shown in Table 2. The total trajectory of dz estimated on the left lower arm is mostly smooth while there are some drastically changes on estimation of dz on right lower leg. The reason for this is because the imprecise estimation on the right upper leg causes a great error on the estimation of right lower leg.

Fig. 17 shows the extraction of Euler angles during the three hundred frames for left lower arm and right lower leg. Part (a) shows the reconstructed trajectory of left lower arm in which we can see that the α angle of it increase from 0° to 360°, part (b) shows the reconstructed trajectory of right lower leg. The picture denotes the left lower arm of subject moves in a circle. From Fig. 17, we note that the reconstructed trajectory of human motion is plausible although it failed to various extents on some frames maybe due to the change of illumination or fault selection of Δz, and the error occurred in the human pose reconstruction of previous frame will not impact that of next frame.

From this experiment, we know that the reconstruction efficiency of the proposed algorithm depends on a number of factors including the accuracy of the change of scale factor s corresponds to a unit change of z, the accuracy with the projections of the joints can be located in the image, and the accuracy of the estimates for the relative length of the segments in the human skeleton model.

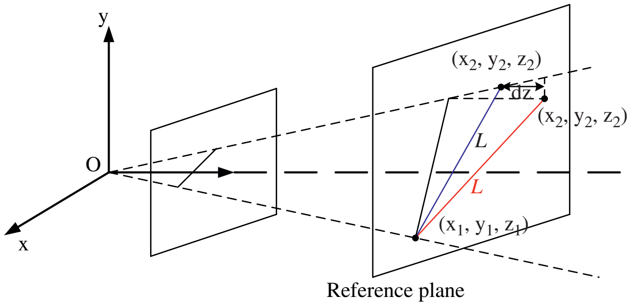

Fig. 18. Different project models.

6.4. Comparison to scaled orthographic projection

In Section 6.2 we have shown that the proposed method to estimate depth value can present encouraging reconstructed results. To prove the efficiency of the proposed method we will compare it to the widely used scaled orthographic projection [16,22]. The accuracy of both methods was compared for a challenging test sequence.

The test sequence (90 frames) shows a person performing full articulation. Some frames are shown in Fig. 19. First, the person hits

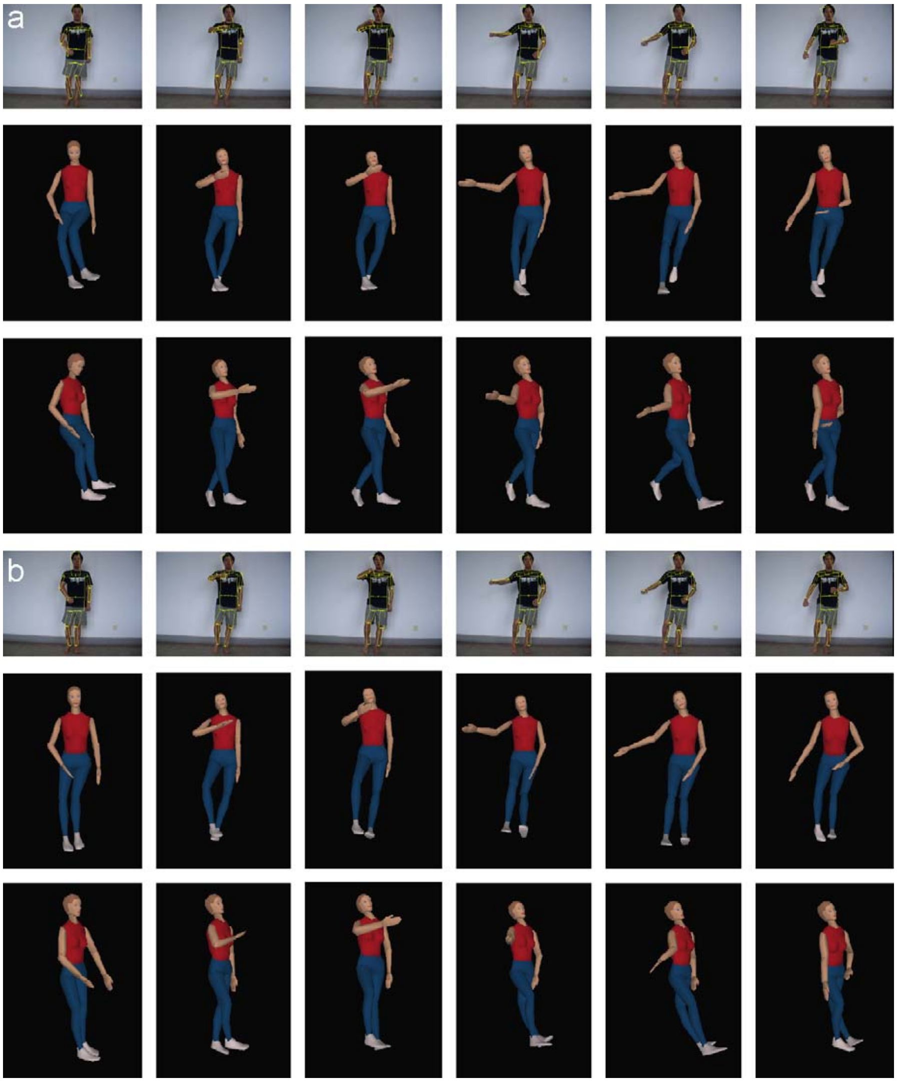

Fig. 19. Comparison between scaled orthographic projection method and the proposed method. The first three rows show the poses for six frames (15, 30, 45, 60, 75, 90) obtained under scaled orthographic projection. The results for the proposed method are shown in the last three rows. (a) Reconstructed results estimated by scaled orthographic projection. (b) Reconstructed results estimated by the proposed method.

with a sweeping motion of the arm and then he kicks with both legs and keeping the arms in a defensive manner close to the upper body. The tracker must be able to keep the arms separated from the torso and should survive the fast motions.

The scale factor, s, used in the scaled orthographic projection was obtained from the scale factor s corresponding to root joint J_1. For our comparison with scaled orthographic projection, the same histogram matching was used. Two human tracking were both performed on the original images.

Our proposed method to estimate depth value was successful in tracking the whole sequences as can be seen for the six selected frames in Fig. 19. The scaled orthographic projection was not able to track the sequences with the same efficiency. The reason for this is that the depth value between two adjacent joints estimated by the scaled orthographic projection is not a ground truth. Fig. 18 demonstrates the differences between scaled orthographic projection and perspective projection, the blue line and the red line represent the reconstructed object under perspective projection and scaled orthographic projection, respectively, dz denotes the difference between the two depth values reconstructed under perspective projection and scaled orthographic projection. Since the depth value estimated by scaled orthographic projection deviate from the ground truth, the estimated Euler angles of joints will deviate from the ground truth accordingly. Therefore, the obtained degree of similarity is inaccurate while the template is projected onto the image plane, and the tracked joints will deviate from the corresponding ground truth ones.

The tracking accuracy was done by comparing the resulting pixels coordinates of both methods to the ground truth data. Table 3 shows the computed average pixel errors for both methods over the test sequences. The average pixel error is defined as

$$\text{The average pixel error} = \frac{\sum_{i=1}^{90}\sqrt{(x_i - x_i')^2 + (y_i - y_i')^2}}{90} \tag{19}$$

where (x_i, y_i) are the coordinates of manually labeled ground truth joints in the ith frame, (x_i', y_i') are the estimated coordinates of joints in image in the ith frame. As shown in Table 3 there is some improvement in results for the proposed method compared to the scaled orthographic projection method.

Table 3
The average pixel error for both methods over the test sequences.

Joint	Average pixel error	
	Scaled orthographic projection	Proposed method
Right elbow	14.3423	10.4453
Right wrist	17.1907	13.5435
Right knee	16.6745	10.1255
Right ankle	20.3454	14.4012
Left elbow	12.6864	8.6904
Left wrist	14.2345	10.4528
Left knee	15.5623	11.4576
Left ankle	23.8068	14.8504

7. Conclusion and future research

We proposed an algorithm to automatically reconstruct 3D human motion pose from uncalibrated monocular video sequences. A key feature of our approach is the proposed method to reconstruct 3D human pose from the corresponding 2D joints on the image plane. In the experiments, the human 3D pose reconstruction is accomplished automatically or manually.

There are several advantages of the proposed approach. First, no camera calibration is needed as required by previous approaches that use multiple-camera setups. Second, the approach exploits invariance of skeleton constrains and morphing appearance model matching to obtain better 3D structure estimates. Finally, the method requires no special constrains to background.

In future, we intend to enhance the flexibility of this algorithm enable it to handle human motion in Z direction displacement. Furthermore, our work will focus on automatic locating the position of joints in the first frame.

Acknowledgements

This research was funded by National Science Foundation of China under Contracts 60673093, 90715043 and 60803024, and supported by Program for Changjiang Scholars and Innovative Research Team in University of China IRT0661.

References

[1] Y. Song, X. Feng, P. Perona, Towards detection of human motion, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2000, pp. 1810–1817.

[2] R. Rosales, M. Siddiqui, J. Alon, S. Sclaroff, Estimating 3D body pose using uncalibrated cameras, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2001, pp. 821–827.

[3] N. Howe, M. Leventon, B. Freeman, Bayesian reconstruction of 3d human motion from single-camera video, in: Proceedings of the Neural Information Processing Systems, 1999, pp. 820–826.

[4] K. Sminchisescu, B. Triggs, Covariance scaled sampling for monocular 3D body tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2001, pp. 447–454.

[5] K. Grauman, G. Shakhnarovich, T. Darrell, Inferring 3D structure with a statistical image-based shape model, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2003, pp. 641–648.

[6] A. Yilmaz, O. Javed, M. Shah, Object tracking: a survey, ACM Computing Surveys 38 (4) (2006) 1–45.

[7] L. Wang, T.L. Tan, Recent developments in human motion analysis, Pattern Recognition 36 (3) (2003) 585–601.

[8] K. Rohr, Incremental recognition of pedestrians from image sequences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1993, pp. 9–13.

[9] C. Bregler, J. Malik, Tracking people with twists and exponential maps, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1998, pp. 8–15.

[10] D.M. Gavrila, L.S. Davis, 3D model-based tracking of humans in action: a multiview approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1996, pp. 73–80.

[11] C. Sminchisescu, B. Triggs, Kinematic jump processes for monocular 3d human tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2003, pp. 69–77.

[12] K. Toyama, A. Blake, Probabilistic tracking with exemplars in a metric space, International Journal of Computer Vision 48 (1) (2002) 9–19.

[13] C. Sminchisescu, B. Triggs, Covariance scaled sampling for monocular 3d body tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2001, pp. 447–454.

[14] J. Deutscher, A. Blake, I. Reid, Articulated body motion capture by annealed particle filtering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2000, pp. 126–133.

[15] C. Sminchisescu, B. Triggs, Building roadmaps of local minima of visual models, in: Proceedings of the European Conference on Computer Vision, 2002, pp. 566 –582.

[16] T. CJ, Reconstruction of articulated objects from point correspondences in a single uncalibrated image, Computer Vision and Image Understanding 80 (8) (2000) 349–363.

[17] F. Remondino, A. Roditakis, 3D reconstruction of human skeleton from single images or monocular video sequences. in: 25th Pattern Recognition Symposium, Lecture Notes in Computer Science, Springer, Berlin, 2003, pp. 100–107.

[18] X.M. Liu, Y.T. Zhuang, Y.H. Pan, Video based human animation technique, in: Proceedings of the 7th ACM International Multimedia Conference, 1999, pp. 353–362.

[19] E. Peng, L. Li, Estimation of human skeleton proportion from 2D uncalibrated monocular data, in: Proceedings of the Computer Animation and Social Agents (CASA 2005), 2005.

[20] J. Zhao, L. Li, C.K. Kwoh, Posture reconstruction and human animation from 2D feature points, Computer Graphics Forum 24 (4) (2005) 759–771.

[21] N.I. Badler, C.B. Phillips, B.L. Webber, Simulating Humans: Computer Graphics Animation and Control, Oxford University Press, New York, 1993.

[22] C. Barron, I.A. Kakadiaris, On the improvement of anthropometry and pose estimation from a single uncalibrated image, Machine Vision and Applications 14 (4) (2003) 229–236.

About the author—BEIJI ZOU received the B.S. degree in computer science from Zhejing University, China, in 1982, received the M.S. degree from Tsinghua University specializing CAD and computer graphics in 1984, and obtained the Ph.D. degree from Hunan University in the field of control theory and control engineering in 2001. He is a professor in the School of Information Science and Engineering, Central South University, China. His research interests include computer graphics, CAD technology and image processing.

About the author—SHU CHEN is a Ph.D. candidate in computer application technology from Central South University, China. His research interests include computer vision, human motion tracking and recognition, animation.

Please cite this article as: B. Zou, et al., Automatic reconstruction of 3D human motion pose from uncalibrated monocular video sequences based on markerless human motion tracking. Pattern Recognition (2009), doi: 10.1016/j.patcog.2008.12.024