IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 19, 2018, accepted December 31, 2018, date of publication January 7, 2019, date of current version January 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2891172

An Image-Constrained Particle Filter for 3D Human Motion Tracking

XIUKAI ZHAO[1,2], LEI LYU[1,2], (Member, IEEE), JINLING ZHANG[3], AND CHEN LYU[1,2]

[1]School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China
[2]Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology, Shandong Normal University, Jinan 250014, China
[3]School of Information, Renmin University of China, Beijing 100872, China

Corresponding author: Lei Lyu (lvbu007@163.com)

This work was supported in part by the Shandong Key Research and Development Program under Grant 2017GSF20105 and Grant 2018GGX101003, in part by the National Natural Science Foundation of China under Grant 61502505, Grant 61472232, Grant 61572299, and Grant 61602286, in part by the Major Program of Shandong Province Natural Science Foundation under Grant ZR2018ZB0419, in part by the Natural Science Foundation of Shandong Province under Grant ZR2016FB13, and in part by the Shandong Province Higher Educational Science and Technology Program under Grant J16LN09.

ABSTRACT Tracking 3D human motion from monocular video sequences has aroused great interest in recent years. Among these human motion tracking methods, the particle filter is considered as an effective approach. However, the current approaches based on particle filter still have some limitation such as many particles are obviously not consistent with the observed image due to they are independent of the image information. In this paper, we present an image-constrained particle filter approach to track 3D human motion from monocular video clips with the assistance of a pre-captured motion library. We propose two novel particle filtering criteria and design a hierarchical likelihood function. The top layer of the function consists of the particle filtering criteria, and the bottom layer consists of the likelihood functions based on image contours and edge features. We remove those particles that do not match the image significantly at the top level, and the remaining particles are evaluated using the underlying likelihood function. The experimental results show that our method can effectively improve the accuracy of motion tracking and constrain the estimation of human body position.

INDEX TERMS 3D human motion tracking, image constraint, particle filter, monocular video.

I. INTRODUCTION

Human motion tracking from monocular video has received increasing attention in recent years due to its applicability to many areas, including intelligent surveillance system, human-computer interaction, sports training, medical rehabilitation and special effects in movie. To precisely detect 3D human motion, some researchers acquire motion parameters from sensors attached on the human body. However, it needs to build professional motion capture environment and the equipment is hugely expensive for commercial use, therefore it is unsuitable for most normal applications. Hence, most people adopt the markerless human body tracking by video. Although the human motion video is easy to obtain, getting sufficient information from video dataset to compute the parameters of body motion properly is also a difficult problem in recent years. With the rapid development of machine learning technology, computer vision including 3D human motion tracking makes a great progress. A large amount of research has been devoted to the study of 3D human motion tracking, but the goal of it is hard to achieve perfect result owing to following reasons. For one thing, the large number of degrees of freedom in human body configurations lead to heavy computations of searching in high-dimensional state space. For another, there are a series of factors remarkably influencing tracking results need to be taken into account, such as loss of depth information, complex image backgrounds, differences in body shape and clothing, occlusion and self-occlusion between the limbs and the torsos, limb confusion, etc.

In virtue of the particle filtering can deal with any form of probability distribution, therefore since the publication of the CONDENSATION algorithm [1], [2], this kind of methods quickly become a mainstream in the field of human motion tracking. Currently, in most of commonly-used particle filter-based approaches, the generation of particle sets is independent of image information, leading to a large number of particles of the set are obviously not consistent with the observed image and wasting a lot of computing resources. As a result, many current existing particle filter-based methods suffer from inaccurate human motion prediction and sensitivity to complex and fast-changing motions.

In this paper, we propose a novel 3D human motion tracking technique from monocular video, namely the Image-Constrained particle filter. It is based on the particle filter framework but adds two selecting criteria, which can be estimated by sample data. Based on this, we design a hierarchical particle likelihood function. When evaluating particles, we first remove the particles that obviously do not match the image according to the top-level criterion, and get a better particle collection which is in good agreement with the image. Then only the particles who are selected can enter the next layer of particle evaluation. Experiment results show that our technique could effectively improve the accuracy of 3D human motion tracking.

The structure of this paper is as follows: we review previous work in Section 2. In Section 3, we give a novel particle selection strategy based on image constraints. On the basis, we describe the entire 3D human motion tracking process by applying our proposed Image-constrained particle filter in Section 4. Section 5 shows experimental results. Finally, we conclude this paper in Section 6.

II. RELATED WORKS

There have been a large number of previous research about 3D human motion tracking in the last twenty years. These works can be divided into two categories: the top-down methods and the bottom-up methods.

In the former kind of methods, the human motion is estimated from image features. Some researchers estimate human posture by geometric constraints. They utilize the hinge structure and limb length constraints of the human body to estimate the corresponding 3D joint coordinates according to the position of the 2D joint point in the image. The representative work is the 3D posture estimation method based on hinge body proposed by Taylor [3], who uses the weak perspective projection as the camera model. Mori and Malik [4] first localize joint positions in 2D and then lift them to 3D using the geometric method of Taylor [3]. Chen and Chai [5] simultaneously reconstructs 3D human motion from a small set of 2D image features tracked from uncalibrated monocular video sequences. Wei and Chai [6] identifies a set of new constraints in 2D image and uses them to eliminate the ambiguity of 3D pose reconstruction. Wei and Chai [7] formulates the video-based motion modeling process in an image-based keyframe animation framework. This type of methods extract useful information from the image to synthesize the human body pose. The process is morbid due to factors such as image noise, limb occlusion, and lack of depth information. The main difficulty is that the mapping from image features to poses is one-to-many.

The other researchers retrieve similar posture from the 3D motion capture data. Howe [8] constructs a motion tracking system Silo Tracking based on contour retrieval. Shakhnarovich *et al.* [9] propose a human motion estimation method based by virtue of motion database index. The retrieve process of the two methods are all based on the silhouette. The advantage is that it is not affected by light and clothing color. But the ability to distinguish is limited, different postures may get very similar contours. Therefore, Poppe [10] adopts the K-Nearest Neighbor (KNN) interpolation method to estimate human posture. And there are also propose that regress posture from image features using functions to approximate human posture in the image or video or fitting such functions by supervised or semi-supervised learning methods. Brand *et al.* [11] use the Hidden Markov Model (HMM) to establish a statistical model for 3D human motion. Howe *et al.* [12] directly train the mapping function from image sequence to posture sequence. Agarwal and Triggs uses the Relevance Vector Machine (RVM) [13] as posture regression model [14], [15]. In [16], Agarwal proposes another regression scheme. He uses a regression method similar to the Mixture of Experts (MoE) [17], [18]. In addition, the probability density propagation model [19], [20] and Local regression model [21] are applied to the human motion tracking. The above methods based on retrieval and regression generally fail to restore the spatial position of the human body because image features (human contours, the texture features, etc.) typically contain only the appearance information of the human body without the positional information of the human body relative to the environment. Moreover, these methods can only recover the pose in the training sample and generally require a large number of labeled samples.

In the latter kind of methods, the 3D human motion is often optimized in a prediction-correction manner. These methods minimize the cost of matching between the projection of the posture and the observed image by adjusting the posture. The core of them is the optimization algorithm. Among them the most famous approach is the particle filtering [22], which is derived from the Bayesian theory and the Monte Carlo method. It can solve the nonlinear problem such as 3D human motion tracking that the classical Kalman filter cannot solve. In which the possible distributions of target's states are depicted by a group of weighted particles [23]. Liu and Payandeh [24] propose an approach for tracking movements of a person based on the notion of a hierarchical particle filter which incorporates two layers consisting of coarse-to-fine tracking subsystems. Sidenbladh *et al.* [25] has successfully used the particle filter to track human motion in short monocular video clips ($<$100 frames). Since the human body has a high dimension, a large number of particles are required to effectively represent the posterior distribution. Doucet [26] and Doucet *et al.* [27] use the Rao-Blackwellisation (RB) technique [28] to reduce the dimension of state space. Xu and Li [29] employ the Partial Least Squares Regression [30] to learn the relationship between the left and right limbs. Other scholars simplify the number of particles by reducing the dimension of state space. Sidenbladh *et al.* [31] and Urtasun *et al.* [32] use the Principal Component Analysis (PCA) to reduce the dimension of motion segments. Urtasun transforms the pose space into a low-dimensional shape using the Gaussian Process Latent Variable Model (GPLVM) and the Gaussian Process Dynamical Model (GPDM) [33], [34] Hidden variable space, then tracked in this space. Sminchisescu and Jepson use the Laplacian Eigenmaps [20] and the Laplacian Eigenmaps Latent Variable Models (LELVM) [35] to learn the mapping relationship from pose space to low-dimensional hidden variable space. Li *et al.* [36] use a Mixture of Factor Analyzers to transform high-dimensional poses into a Globally Coordinated Local Linear Models. Liu *et al.* [37] propose an exemplar-based conditional particle filter (EC-PF) from monocular camera by introducing a conditional term with respect to exemplars and image data. Chang and Lin [38] propose a novel progressive particle filter comprises three principal techniques: hierarchical searching, multiple predictions, and iterative mode-seeking. Bao *et al.* [39] present a particle filter based human position estimation method using a foot-mounted inertial and magnetic sensor module. Du *et al.* [40] present a differential evolution-Markov chain (DE-MC) particle filtering by taking the advantage of the DE-MC algorithm's ability to approximate complicated distributions.

In these methods, the human pose is obtained by optimizing the cost function between the projected contour and the contour of the image. Since the dimension of the human body posture is relatively high, the search process is in a high-dimensional posture space. Moreover, the cost function is very complex, making the searching more difficult. On the whole, the top-down approaches are capable of restoring body position and orientation information and even can restore arbitrary poses theoretically, due to they do not rely on any training samples.

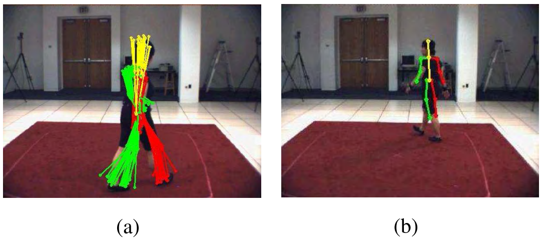


FIGURE 1. The phenomenon that the particles do not match the contour of the image.(a) Some joints are projected outside the contour of the image. (b) The projection of particles cannot cover the outline of the human body.

III. PARTICLE SELECTION STRATEGY BASED ON IMAGE CONSTRAINT

In the process of 3D human motion tracking based on the particle filtering, there is often a mismatch between the particle and the contour of the image, resulting the joint points obtained after the particle projection are outside the contour of the human body (see Figure 1(a)). These mismatched particles will affect the accuracy of posture estimation, so it is necessary to distinguish these particles and delete them from the particle collection. The common method of deletion is to preserve only the particles that are projected entirely inside the contour of the human body. But this will result the projection contour is significantly smaller than the image contour (see Figure 1(b)) and ultimately affect the accuracy of human motion tracking. As a result, we design two novel criteria for the particle selection.

A. PARTICLE SELECTING CRITERIA

Criterion I: Limb constraint - the limb projection is inside the outline of the image.

Assuming that the contours of the human body have been extracted relatively clearly, the 2D limbs obtained from all particle projections must be located inside the contour of the human body. Considering the influence of contour noise, we performed a small expansion process on the contour of the human body in the image. In order to simplify the calculation, we separately sample the projections of the four limbs to obtain discrete projection points, as shown the green points in Figure 2. If all of these points are within the outline, it indicates that these particle satisfies **Criterion I**, otherwise they particle is will be deleted.

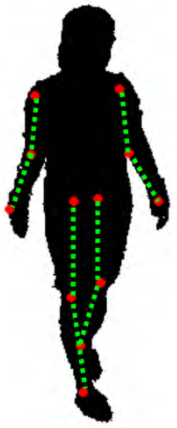


FIGURE 2. Limb constraint.

Criterion II: Joint constraint - the projections of the head and foot joints are within the limits of the outline of the image.

Though using **Criterion I** can guarantee all joint projections are inside the outline of the image, the projection of human body may shrink into a certain part of the image contour, which results in the enlarged positional deviation of the human body. Therefore, we need to limit the size of the projection contour of human body so that it can cover the image contour as far as possible. Due to the high complexity of computing body projection contour, we use the human body joint constraints to approximate this limitation. That is, the projection of the head joint and the ankle joint must be within the limits of the image contour. By this way, we can obtain the effective particles, as shown in Figure 3.

B. LEARNING THE CRITERIA

For **Criterion I**, we manually set up an expansion coefficient. In our experiment, all contours are expanded by 5 pixels.

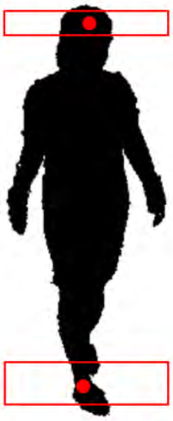


FIGURE 3. Joint constraint.

The mathematical description of the **Criterion I** is expressed as Equation 1.

$$P_{sil}(I|x) \propto \delta(Proj(x)\ in\ Sil(I)) \quad (1)$$

Here the $Proj(x)$ is the projection of the main limbs of the posture x, the $Sil(I)$ denotes the human contour in the image, and the $\delta(x)$ is expressed as:

$$\delta(x) = \begin{cases} 1, & x = ture \\ 0, & x = false \end{cases} \quad (2)$$

The value of the $\delta(x)$ is 1 when the limb projection of the pose x is within the image contour, otherwise is 0.

For **Criterion II**, we extract the relative position of the head and ankle joints in the image contour from the training set.

In order to obtain the prior distribution, we project the 3D human pose in the training set onto the 2D image plane and calculate the relative position of the head and ankle joint in the bounding box of the image contour. Since the feet are lifted alternately with one part of a foot always on the ground during walking, the ankle of the foot which touches the ground is at the lower position of the feet, and the other foot moves in a wide range. Therefore we use the coordinates of the ankle joint at the lower position to denote the 2D position distribution of the ankle.

To make the coordinates are independent of the size of the human body contour, we normalize the coordinates of the two joints into the bounding box of the human body contour and the obtained coordinates are in the range of [0, 1]. Figure 4 shows the 2D coordinate distribution of the head and ankle joints. From which, we can find that the variation range of x-coordinate of the two joints is basically within [0.1, 0.9], and the variation range of y-coordinate is extremely small. Therefore, the y-coordinate can be used to constrain particles. We define the prior distribution of the y-coordinate of the two joints by Equation 3.

$$\begin{cases} P_{head}(y) \propto \delta(l_{head} \leq y \leq \mu_{head}) \\ P_{foot}(y) \propto \delta(l_{foot} \leq y \leq \mu_{foot}) \end{cases} \quad (3)$$

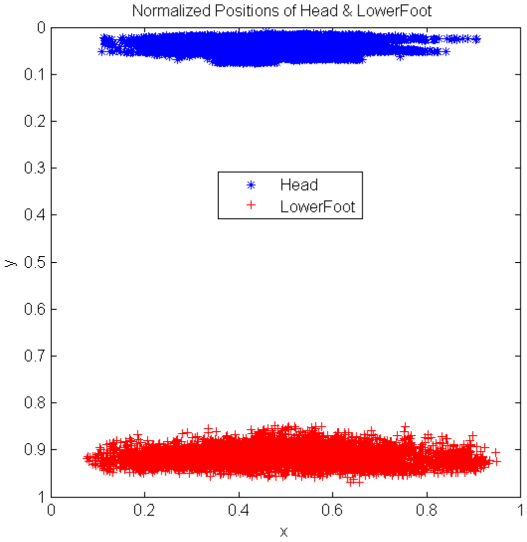


FIGURE 4. 2D coordinate distribution of the head and ankle joints.

In Equation 3, we use training samples to estimate the parameters l_{head}, μ_{head}, l_{foot}, μ_{foot}.

IV. 3D HUMAN MOTION TRACKING

In our method, we first prune some particles according to the two criteria mentioned above before calculating the weight of particles, only the remaining particles can be performed the particle evaluation operation. The corresponding particle likelihood function is described by Equation 4.

$$P(I_t|x_t) = exp(-aE_{sil} - bE_{edge})P_{sil}(I_t|x_t)P_{foot}(x_t)P_{head}(x_t) \quad (4)$$

It is a hierarchical likelihood function and can be divided into two layers. The first layer is the selecting criterion $P_{sil}(I_t|x_t)P_{foot}(x_t)P_{head}(x_t)$, and the second one is the likelihood function based on contour and edge features $exp(-aE_{sil} - bE_{edge})$. We firstly perform the particle evaluation on the first level and select the particles whose weight is 1 based on the evaluation results. Then these selected particles are carried out the likelihood function estimation. The benefit of this is those particles that do not meet the constraints can be removed and the remaining particles have a chance to get more complex evaluations.

A. PARTICLE EVALUATION FUNCTION

In this part, we design the particle evaluation function using the image contours and edge features.

1) CONTOUR FEATURES

We use the background subtraction method to extract the contour of human body in the video by means of the background subtraction method. The matching degree between the projection contour and the image contour (Figure 5) is represented by the bi-directional silhouette matching [41],

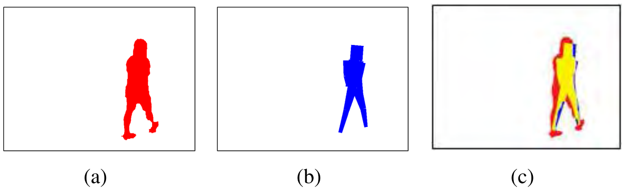


FIGURE 5. Sketch of contour matching cost. (a) Contour of human body. (b) Contour of model projection. (c) Sketch of contour matching cost.

which is calculated as follows:

$$E_{sil} = (1-\alpha)\frac{B}{B+Y} + \alpha\frac{R}{R+Y} \tag{5}$$

In this formula, the R denotes the area of the body contour in the image that is not covered by the projection contour of the model (Figure 5), corresponding to the red part in Figure 5(c); the B represents the area of the projection contour of the model that is not covered by the image contour, corresponding to the blue part in Figure 5(c); the Y is the area of the overlap of two contours, corresponding to the yellow part in Figure 5(c). The value of the E_{sil} is 0 when the two contours completely overlap, conversely its value is 1 when there is no overlap between them.

Particularly, the R, B and Y are defined in Equation 6:

$$\begin{cases} R = \sum_p M^f(p)(1-M^b(p)) \\ B = \sum_p M^b(p)(1-M^f(p)) \\ Y = \sum_p M^f(p)M^b(p)) \end{cases} \tag{6}$$

where the M^f and M^b respectively express the body contour of the image (Figure 5(a)) and the projection contour of the model (Figure 5(b)), the p denotes the pixel index of an image, the α is an adjustment parameter and its value is 0.5 in our experiment.

2) THE FEATURE OF EDGE

We use the Canny edge detection operator to extract the edge of the image, and the resulting edge graph contains many edges of the background (Figure 6(c)). Since the background edges may interfere with the evaluation of particles, so we use the human body contour (Figure 6(b)) to suppress them. Let the f represent the foreground contour map, we expand the f by about 5 pixels to get a new contour map g. We keep the edge map corresponding to the foreground part of g, and get the suppressed edge (Figure 6(c)). This edge map basically contains the edges on the contour of the human body, and the edges in the background are effectively suppressed.

The evaluation function of particle edge matching is computed by the truncated chamfer distance (*TCD*) [42] between the edge of image and the edge of model projection contour. Suppose the points set of the model is represented as $U = \{u_i\}_{i=1}^n$ and that of the image is $V = \{v_i\}_{i=1}^m$, then the E_{edge} is calculated as follows:

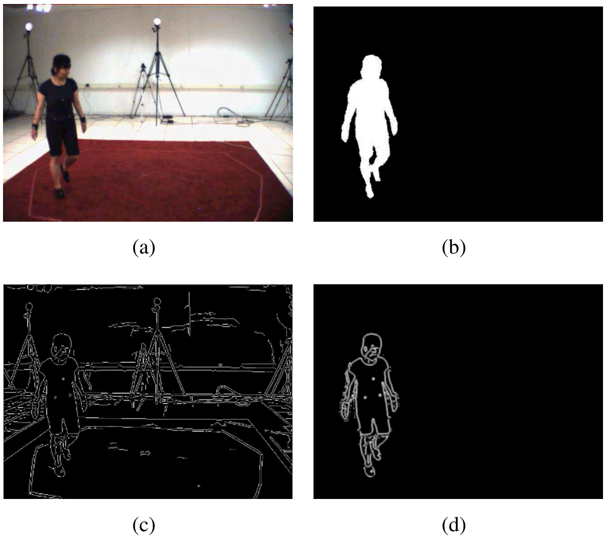


FIGURE 6. Sketch map of human body contour edge extraction. (a) The original image. (b) The contour map. (c) Canny edge map. (d) The edge map after background suppression.

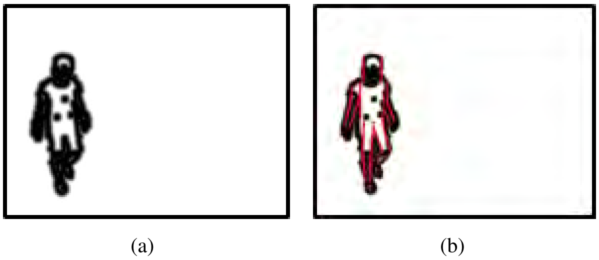


FIGURE 7. Sketch map of edge matching. (a) Map of distance transform of image edge. (b) Sketch map of edge matching.

$$E_{edge} = tcd(U,V) = \frac{1}{n}\sum_i min\left\{min_j \left\|u_i - v_j\right\|, \tau\right\} \tag{7}$$

where τ is a cut-off factor for reducing the influence of outer point (outliers) and missing edges.

We use the Distance Transform (DT) to approximate the *TCD*, then the E_{edge} can be rewritten as:

$$E_{edge} = tcd(U,V) = \frac{1}{n}\sum_{i=1}^{n} min(DT(u_i), \tau) \tag{8}$$

Combining the two matching costs E_{sil} and E_{edge}, the particle evaluation function can be defined as Equation 9.

$$w = P(I|x) = exp(-aE_{sil} - bE_{edge}) \tag{9}$$

Here, the coefficients a, b are used to adjust the weights of different items.

B. 3D HUMAN MODEL AND POSTURE

In this paper, we adopt a simplified 3D human body model, which consists of 10 rigid bodies, as shown in Figure 8.

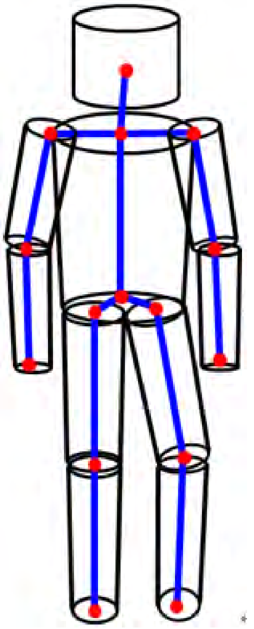


FIGURE 8. **3D human body model.**

The trunk, limbs and head of the model are respectively represented by the elliptical cylinder, truncated cone and cylinder and they are linked by 14 joints.

Let the x denote the 3D human posture. It is composed of three parts and represented as:

$$x = (r, o, p) \tag{10}$$

where, the r is the position of the human body, the o is the rotation of the human body, and the p expresses the coordinates of 14 joints in the local coordinate system of the human body:

$$p = (p_1, p_2, \cdots, p_{14}) \tag{11}$$

Assuming the r, o and p are independent of each other, then the $P(x_t|x_{t-1})$ can be decomposed into three independent dynamic models:

$$P(x_t|x_{t-1}) = P(r_t|r_{t-1})P(o_t|o_{t-1})P(p_t|p_{t-1}) \tag{12}$$

It is proved as follows:

$$\begin{aligned} P(x_t|x_{t-1}) &= P(r_t|r_{t-1})P(o_t|o_{t-1})P(p_t|p_{t-1}) \\ &= \frac{P(r_t, o_t, p_t, r_{t-1}, o_{t-1}, p_{t-1})}{P(r_{t-1}), P(o_{t-1}), P(p_{t-1})} \\ &= \frac{P(r_t, r_{t-1})P(o_t, o_{t-1})P(p_t, p_{t-1})}{P(r_{t-1})P(o_{t-1})P(p_{t-1})} \\ &= P(r_t|r_{t-1})P(o_t|o_{t-1})P(p_t|p_{t-1}) \end{aligned} \tag{13}$$

Because the dimension of state vector x is very high, it need a large number of particles to gain a better tracking result, thus it is time-consuming to evaluate these particles. Essentially the intrinsic dimension of human motion is relatively low, so we use the principal component analysis (PCA) to reduce the dimension of the p.

Given N training samples $\{p_i = (p_{i,1}, p_{i,2}, \cdots, p_{i,14})\}_{i=1}^{N}$, we calculate the eigenvector of the sample covariance matrix $\{v_k\}_{k=1}^{42}$ and the corresponding eigenvalue $\{\alpha_k\}_{k=1}^{42}$. Suppose the $\alpha_k > \alpha_{k+1}$ $(k = 1 \cdots 41)$, we select the first d eigenvectors and project the p_i onto these d eigenvectors, then the p_i can be represented as:

$$p_i \cong v_0 + \sum_{k=1}^{d} g_{i,k} v_k \tag{14}$$

where,

$$v_0 = \frac{1}{N} \sum_{i=1}^{N} p_i \tag{15}$$

We use the $g_i = (g_{i,1}, g_{i,2}, \cdots, g_{i,d}) \in R^d$ to approximate the p_i. And we choose the d to make the subspace $\{v_k\}_{k=1}^{d}$ retain more than 95 of the variance in the original space. For walking motion, the range of D is [3, 10]. In order to make the posture estimation more reasonable, we set the range of the g_i by Equation 16.

$$-3\sqrt{a_m} < g_{i,m} < 3\sqrt{a_m}, \quad m = 1 \ldots d \tag{16}$$

If the $g_{i,m}$ beyond this range, we set it as the nearest boundary value. As a result, we can track the human posture in a low dimensional subspace, then the Equation 12 can be rewritten as:

$$P(x_t|x_{t-1}) = P(r_t|r_{t-1})P(o_t|o_{t-1})P(g_t|g_{t-1}) \tag{17}$$

C. DYNAMIC PROCESS

1) THE DYNAMIC PROCESS OF THE BODY POSITION

Assuming the three components of 3D coordinates of the human body are independent of each other during movement, the dynamic process of human position can be defined as:

$$P(r_t|r_{t-1}) = P(r_{x,t}|r_{x,t-1})P(r_{y,t}|r_{y,t-1})P(r_{z,t}|r_{z,t-1}) \tag{18}$$

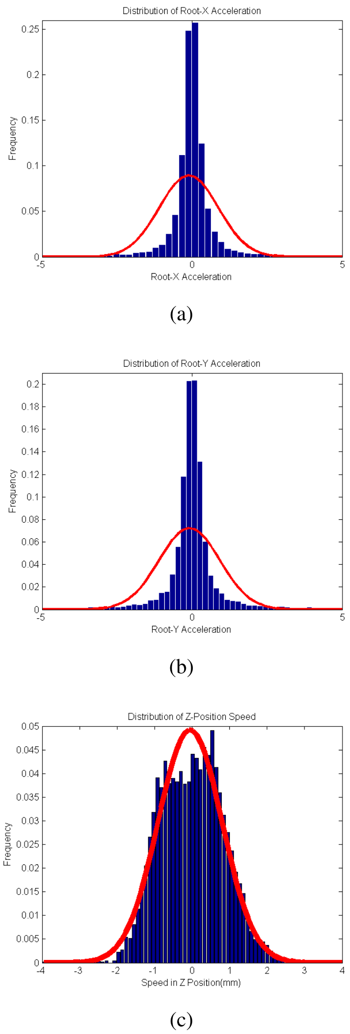
where the $r_{x,t}$, $r_{y,t}$, $r_{z,t}$ respectively represent the x, y, z component of the human body position in time t.

Figure 9(a) and Figure 9(b) show the acceleration distribution of the r_x and r_y, we can find that the acceleration in the two directions can be described by the Gaussian distributions with zero mean. Therefore, it is reasonable to assume that the human motion in the direction X and Y (horizontal plane) is uniform. As a result, we define the dynamic process of the two directions as:

$$\begin{cases} P(r_{x,t}|r_{x,t-1}) = N(r_{x,t}; 2r_{x,t-1} - r_{x,t-2}, \sigma_{rx}^2) \\ P(r_{y,t}|r_{y,t-1}) = N(r_{y,t}; 2r_{y,t-1} - r_{y,t-2}, \sigma_{ry}^2) \end{cases} \tag{19}$$

where the σ_{rx} and the σ_{ry} denote the standard deviation and they can be estimated from training samples.

However, the above dynamic process defined as in Equation 19 is a second-order Gaussian autoregressive process, which contradicts the hypothesis of particle filter. For this reason, we transform the second-order process into a first-order one [43] by replacing $r_{x,t}$ and $r_{y,t}$ with $\tilde{r}_{x,t} = \binom{r_{x,t}}{r_{x,t-1}}$



(a)

(b)

(c)

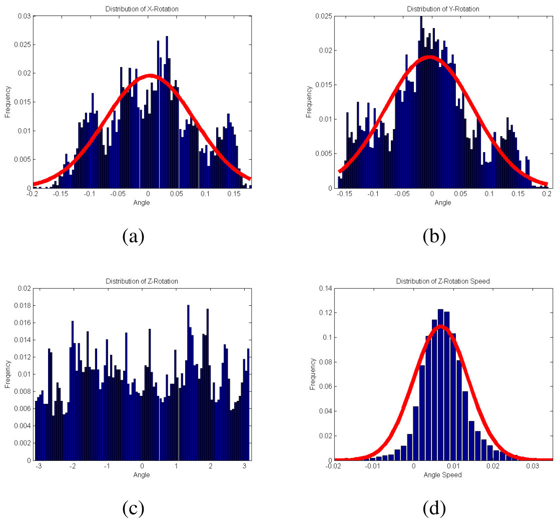
FIGURE 9. **Distribution law of human body position. (a) Distribution of X direction's acceleration. (b) Distribution of Y direction's acceleration. (c) Distribution of Z direction's acceleration.**

and $\tilde{r}_{y,t} = \binom{r_{y,t}}{r_{y,t-1}}$ respectively, then the Equation 19 can be rewritten as:

$$\begin{cases} P(\tilde{r}_{x,t}|\tilde{r}_{x,t-1}) = N(\tilde{r}_{x,t}; \begin{pmatrix} 2 & -1 \\ 1 & 0 \end{pmatrix} \tilde{r}_{x,t-1}, \begin{pmatrix} \sigma_{rx}^2 & 0 \\ 0 & 1 \end{pmatrix}) \\ P(\tilde{r}_{y,t}|\tilde{r}_{y,t-1}) = N(\tilde{r}_{y,t}; \begin{pmatrix} 2 & -1 \\ 1 & 0 \end{pmatrix} \tilde{r}_{y,t-1}, \begin{pmatrix} \sigma_{ry}^2 & 0 \\ 0 & 1 \end{pmatrix}) \end{cases} \quad (20)$$

Since the human body does not move in the vertical direction, the $r_{z,t}$ oscillates in a small range, which is mainly determined by body shapes. Figure 9 shows the velocity distribution of the human body in the Z direction, we can discovery that the position change in the Z direction can also be represented by a Gaussian distribution with zero mean (the red line in the Figure 9(c)), so we establish the dynamic process of the $r_{z,t}$ as following:

$$P(r_{z,t}|r_{z,t-1}) = N((r_{z,t}; r_{z,t-1}), \sigma_{rz}^2) \quad (21)$$



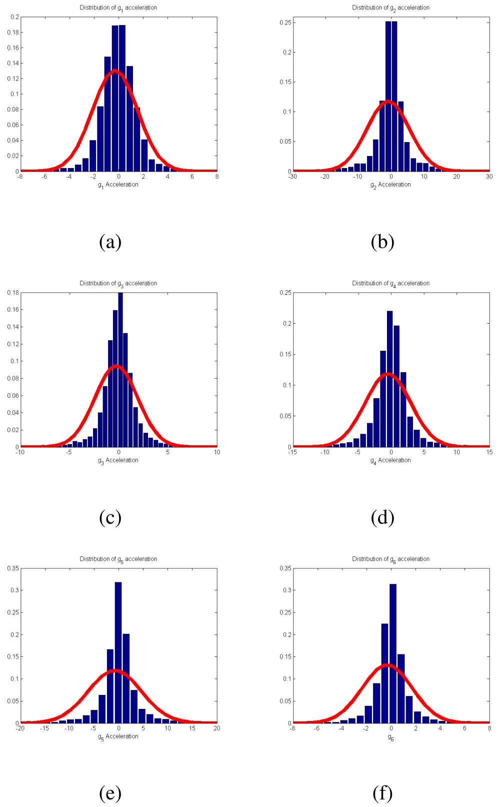
(a)

(b)

(c)

(d)

FIGURE 10. **The distribution law of human body orientation. (a) Distribution of X-orientation. (b) Distribution of Y-orientation. (c) Distribution of Z-orientation. (d) Velocity distribution of Z-orientation.**



(a)

(b)

(c)

(d)

(e)

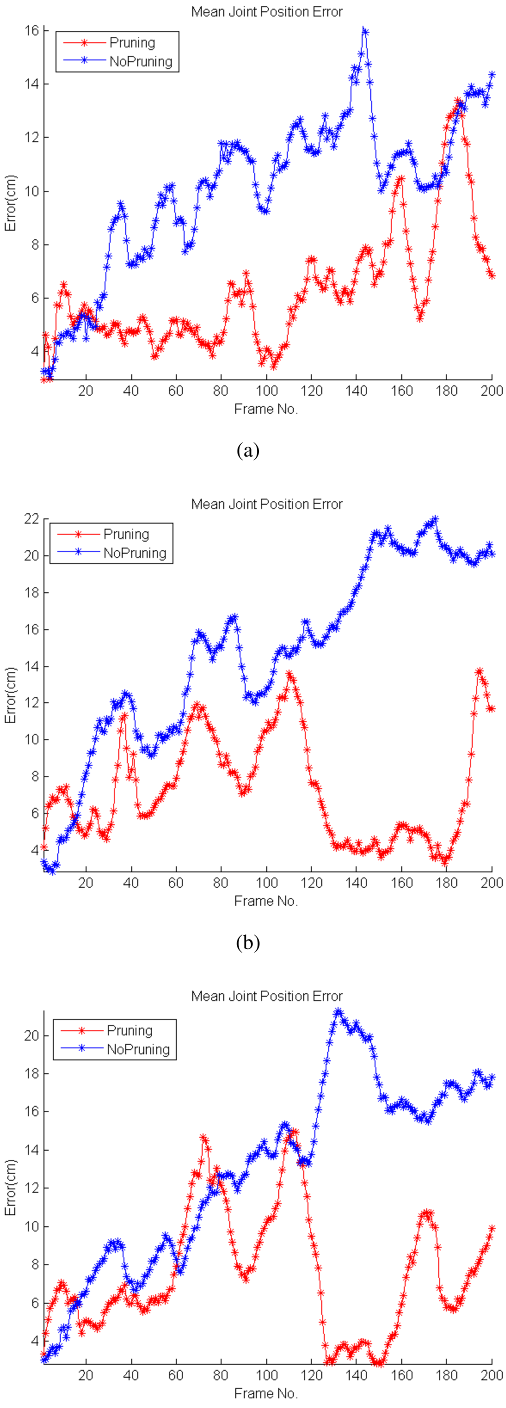
(f)

FIGURE 11. **The distribution law of human posture acceleration. (a) The distribution of the acceleration of g_1. (b) The distribution of the acceleration of g_2. (c) The distribution of the acceleration of g_3. (d) The distribution of the acceleration of g_4. (e) The distribution of the acceleration of g_5. (f) The distribution of the acceleration of g_6.**

In which the standard deviation σ_{rz}^2 is estimated from the sample data.

2) DYNAMIC PROCESS OF HUMAN BODY ORIENTATION

We assume the three components $o_{x,t}, o_{y,t}, o_{z,t}$ of the rotation o_t are independent of each other during the movement, then

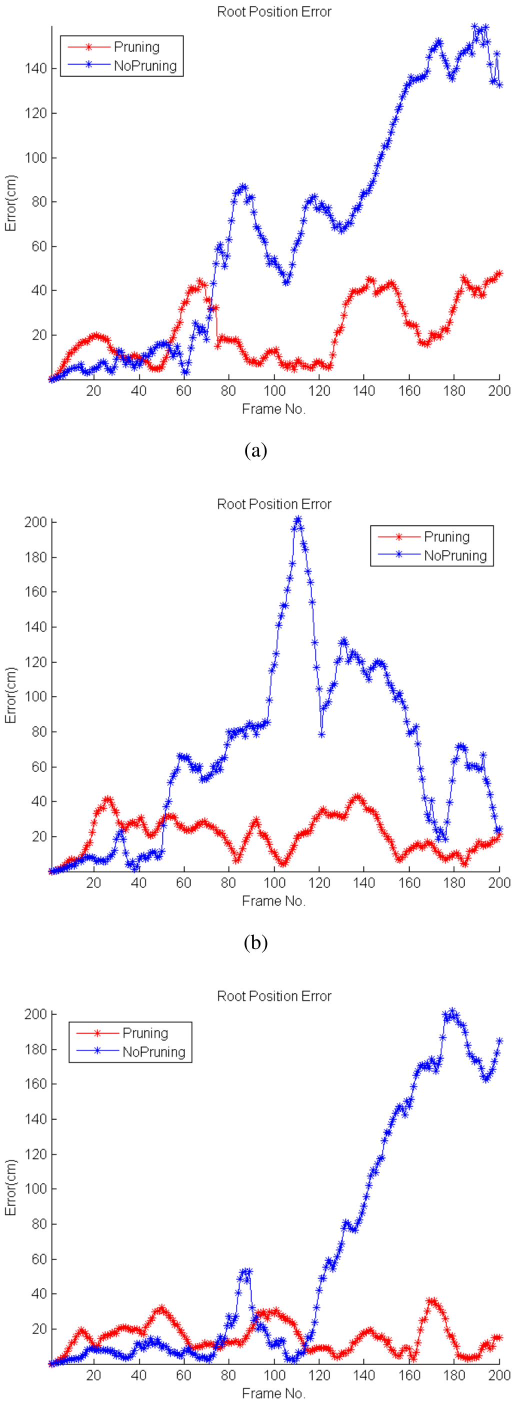


(a)

(b)

(c)

FIGURE 12. Comparison of the joint error. (a) C1. (b) C2. (c) C3.



(a)

(b)

(c)

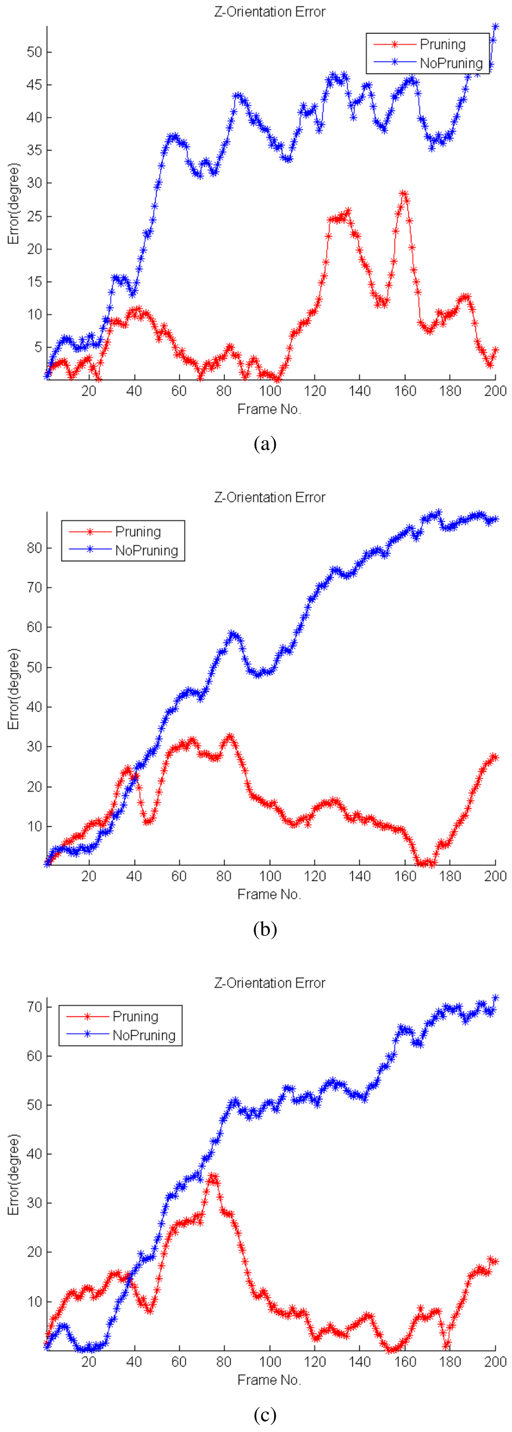
FIGURE 13. Comparison of the position error. (a) C1. (b) C2. (c) C3.

we can define the dynamic process of the o_t as:

$$P(o_t|o_{t-1})=P(o_{x,t}|o_{x,t-1})P(o_{y,t}|o_{y,t-1})P(o_{z,t}|o_{z,t-1}) \quad (22)$$

Figure 10 shows the distribution characteristics of the three orientation components of the human motion. We can see that the o_x and o_y vary in a narrow range, while the o_z varies significantly around the Z axis(the vertical direction).

We use the Gaussian distribution to describe the distribution characteristics of the three components. Since the human motion is a circular motion in space, the o_z is uniformly distributed over the whole $[-\pi, \pi]$ (Figure 10(c)), but the velocity variation of the o_z presents a single peak distribution.



(a)

(b)

(c)

FIGURE 14. Comparison of rotation error. (a) C1. (b) C2. (c) C3.

Based on the above analyses, the whole dynamic process can be described as:

$$\begin{cases} P(o_{x,t}|o_{x,t-1}) = N(o_{x,t};\, \mu_{ox}, \sigma_{ox}^2) \\ P(o_{y,t}|o_{y,t-1}) = N(o_{y,t};\, \mu_{oy}, \sigma_{oy}^2) \\ P(o_{z,t}|o_{z,t-1}) = N(o_{z,t};\, \mu_{oz} + o_{z,t-1}, \sigma_{oz}^2) \end{cases} \tag{23}$$

The parameters μ_{ox}, μ_{oy}, μ_{oz}, σ_{ox}, σ_{oy}, σ_{oz} are estimated from the sample library.

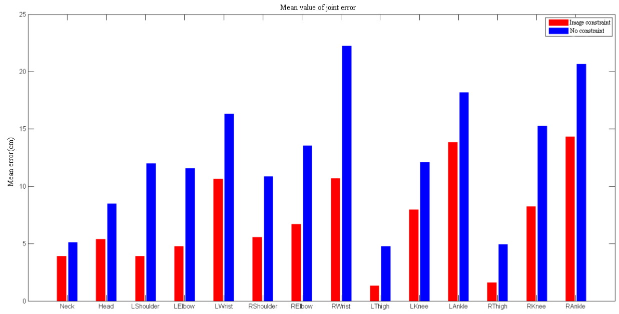


FIGURE 15. Comparison of the mean of the joint error.

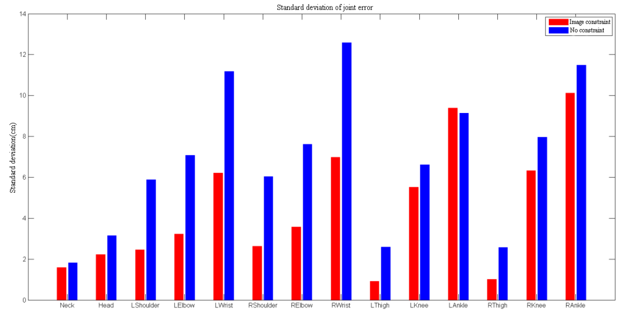


FIGURE 16. Comparison of the standard deviation of the joint error.

In addition, we can find the value of the μ_{oz} (velocity mean of o_z) is greater than 0 from Figure 10(d), this is because the human body is moving in an anticlockwise direction, thus the o_z almost always incremental.

3) DYNAMIC PROCESS OF JOINT COORDINATES

Because the projection from 3D human joint coordinates p_t to a six dimensional subspace can retain more than 95% of the variance, therefore the g_t can be defined as:

$$g_t = (g_{1,t}, \ldots, g_{6,t}) \tag{24}$$

We analyze the distribution characteristics of acceleration in the low dimensional space and discover that the acceleration in all six dimensions presents a sharp unimodal distribution, as shown in Figure 11. Therefore, we use the Gaussian distribution to depict the acceleration distribution of the six dimensions. The mean values of the six accelerations are almost all zero which indicates that the change of the g_t is nearly uniform, that is $g_{t+1} - g_t \approx g_t - g_{t-1}$. Hence, the dynamic process of the g_t can be represented by the uniform motion.

Assuming all dimensions of the g_t are independent, the $P(g_t|g_{t-1})$ can be expressed by Equation 25.

$$P(g_t|g_{t-1}) = \prod_{k=1}^{6} P(g_{k,t}|g_{k,t-1}) \tag{25}$$

Based on the assumption of the uniform motion, the $P(g_{k,t}|g_{k,t-1})$ is defined as:

$$P(g_{k,t}|g_{k,t-1}) = N(g_{k,t};\, 2g_{k,t-1} - g_{k,t-2}, \sigma_{gk}^2) \tag{26}$$

where the σ_{gk}^2 can be trained from samples.

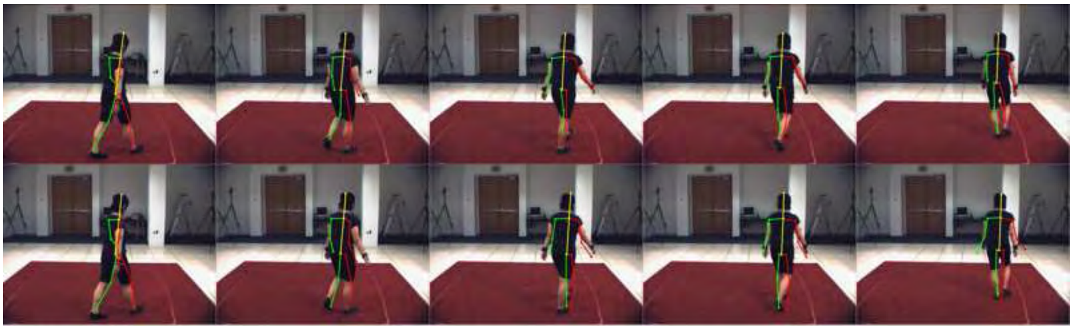


FIGURE 17. Comparison of the tracking results on C1.

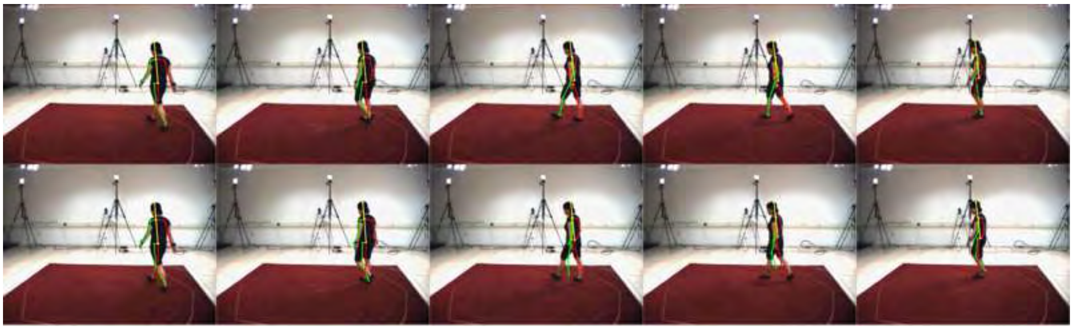


FIGURE 18. Comparison of the tracking results on C2.

By combining all the analysis results, we get the dynamic process model of human body posture. The model is composed of multiple simple products of one-dimensional Gaussian distribution, by which we can easily sample the distribution.

V. RESULTS

A. DATA PREPARATION

We test our approach based on the HumanEvaI dataset provided by Brown University [41]. The HumanEvaI uses the ViconPeak (motion capture device) to collect 3D human motion data while gathering corresponding video data and contains six different movements of four capture objects (S1, S2, S3, S4). Except the object S4, the same action of other object is acquired three times (Trial1-Trial3). At the same time, the synchronized video data is collected from seven different perspectives with three color cameras (C1, C2, C3) and four grayscale cameras. The Trial1 includes the synchronized video data and the 3D motion data, while the Trial2 and Trial3 only contain the video data and the 3D motion data respectively.

We track the gait motion in HumanEvaI, using Trial3 as training data to estimate the parameters of the dynamic model, and the Trial1 as test data.

B. EVALUATION INDEX

In this paper the human posture is represented as the position of the root joint, the orientation of each joint and the coordinates of each joint. We separately measure the error of the three parts. Due to the coordinate of each joint is represented by the offset relative to the root joint, thus the calculated error does not include the human body position error but includes the orientation error. In addition, the errors of the position, rotation and joint are respectively expressed by the Euclidean distance, the absolute value of the deviation of the three angles and the average Euclidean distance between joint coordinates.

C. EXPERIMENTAL RESULTS

1) IMAGE CONSTRAINT EFFECT

In this paper, we perform a comparison of the joint error, human position error and z-axis orientation error based on the three test videos. We can see that the three kinds of errors have been greatly reduced by our particle selection method (Figure 12, Figure 13 and Figure 14). The most obvious effect of the algorithm is the human body position is well limited (see Figure 13), and the difference between the calculated position and the real position is no more than 50 cm. In addition, we can find the errors of human joints and the standard deviations are significantly reduced after particle selecting from Figure 15 and Figure 16.

We also perform a comparison of video tracking results based on the dataset C1-C3, as shown in Figure 17-Figure 19. The first line of the three figures presents the tracking result of our approach, and the second line demonstrates the tracking result of classical particle filter. We can see that the result of

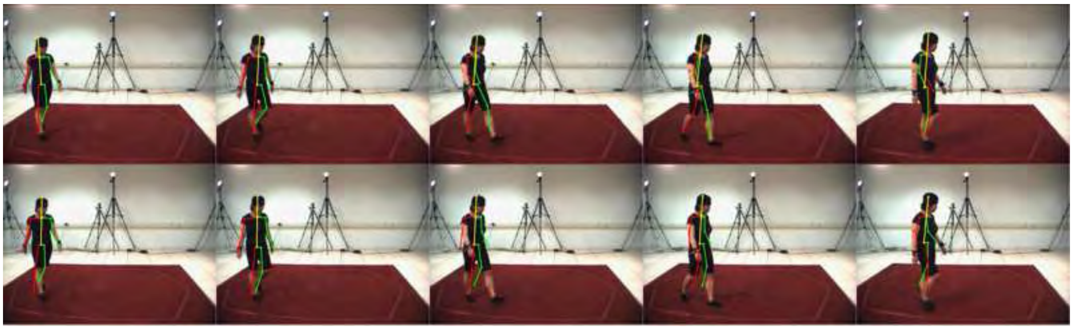


FIGURE 19. Comparison of the tracking results on C3.

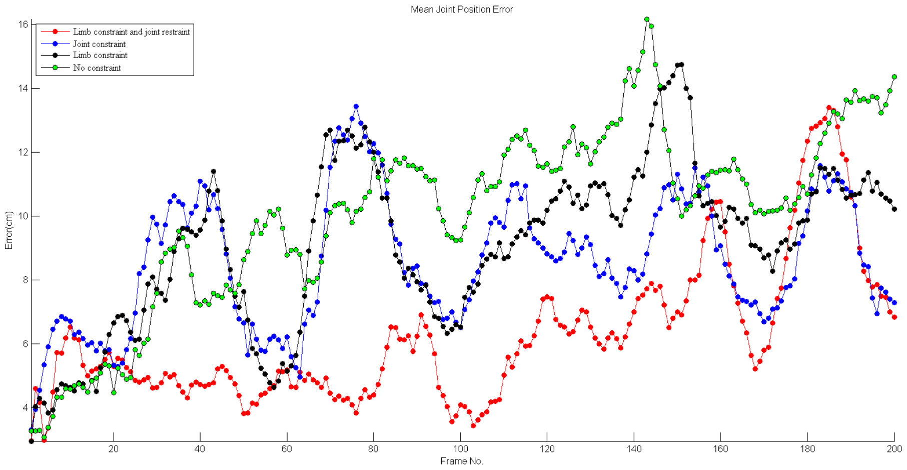


FIGURE 20. Comparison of the mean of joint errors.

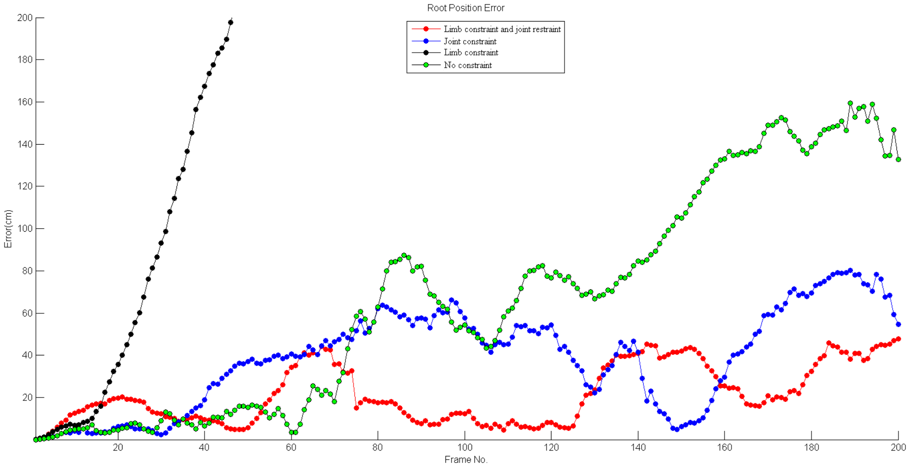


FIGURE 21. Comparison of the body position errors.

our method is much better than that of the classical particle filter.

2) COMPARISON OF DIFFERENT CONSTRAINTS

To further verify the effectiveness of our proposed method, we carry out the experiment on the video C1 including different constraints and without constraints (see Figure 20-Figure 22). Experiment results show that the errors of the joint, position and orientation can be effectively reduced by using the two constraints simultaneously, whereas the position of human body cannot be limited by using the limb constraint alone. This is because it causes the projection

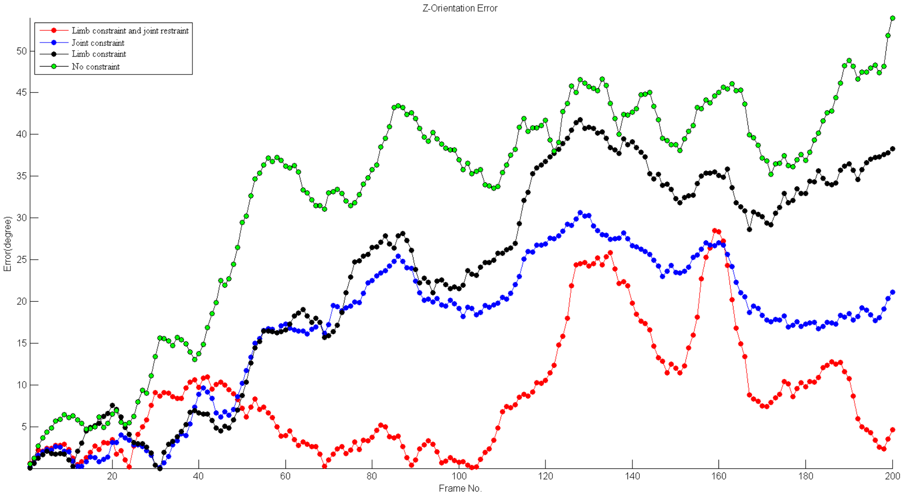


FIGURE 22. Comparison of the human orientation error.

TABLE 1. Mean and standard deviation of joint error.

Dataset	Our method	CONDENSATION	APSOPF
C1	3.36(2.52)	5.05(5.84)	6.23(5.67)
C2	4.37(3.21)	7.13(7.39)	8.35(6.34)
C3	3.01(2.03)	6.24(5.77)	7.87(7.71)

TABLE 2. Mean and standard deviation of maximum joint error per frame.

Dataset	Our method	CONDENSATION	APSOPF
C1	10.08(4.23)	15.55(10.56)	18.09(6.71)
C2	10.16(3.48)	22.11(11.64)	20.13(6.62)
C3	9.96 (4.89)	16.29(8.28)	22.63(12.24)

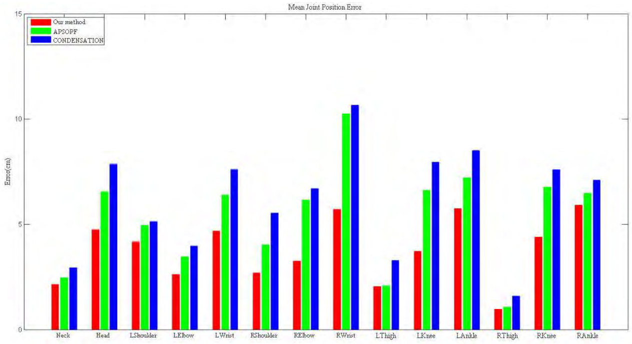


FIGURE 23. Comparison of the mean error of each joint with the three methods.

of the human body to be surrounded by the contour of the image as much as possible, resulting in the estimated position of the human body gradually away from the camera.

3) COMPARISON WITH OTHER METHOD

We compare the tracking results based on our Image-Constrained particle filter with those obtained by implementing the CONDENSATION and APSOPF methods. The experiment is performed on the three video sequences C1, C2, C3, Table 1 and Table 2 show the average error of the joint coordinate and the average maximum error of all joints at each frame, respectively. As can be seen from these two tables, our approach achieves even better performance than the other two methods. We also compare the mean error and standard deviation of each joint using the three tracking methods. As shown in Figure 23 and Figure 24, the errors of almost all joint coordinates obtained by our technique are significantly lower than those of the other two methods.

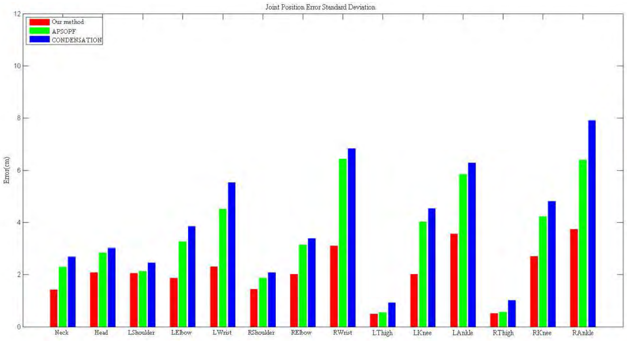


FIGURE 24. Comparison of the standard deviation of each joint with the three methods.

VI. CONCLUSIONS

In this paper, we propose a novel particle selection strategy based on the image constraints. It is composed of two novel selecting criteria, which can be estimated by samples. Their computational complexity is lower than that of particle evaluation. Based on the strategy, we design a hierarchical particle likelihood function. We first conduct the particle evaluation at the top level. The particles that do not meet these two criteria will be deleted, and the remaining particles will enter the next level of likelihood function evaluation. After selecting, some unreasonable particles can be avoided to participate in the evaluation and reduce computation amount. More seriously, these unreasonable particles are likely to gain higher weights in the particle evaluation stage, which will lead to a large difference between the posterior distribution and the real distribution of the attitude represented by the particle set, leading to a large deviation in 3D human motion estimation. Whereas, this can be effectively reduced by our approach. Experiment results show that our technique can effectively improve the accuracy of 3D human motion tracking and constrain the estimation of human body position at the same time. In the light of the method of top-down, we put forward the concept of 3D human motion tracking by the Image-Constrained particle filter, on account of our method can estimate the human body's position in space and does not need the image samples of the annotated posture.

However, our method has several disadvantages. Firstly, both of the criteria depend on image contour, so the human body contour in video is required to be extracted clearly. In addition, we find that the number of particles after particle selecting is sometimes very small, and thus it is difficult to estimate the reliable posterior distribution. In view of this situation, we need to re-sample the prior distribution until there are enough particles to calculate the next likelihood function.

With regard to future research, most of our attention will be focused on extracting higher level information from images, such as the position of human limbs, even the 2D human posture.

REFERENCES

[1] M. Isard and A. Blake, "Contour tracking by stochastic propagation of conditional density," in *Proc. Eur. Conf. Comput. Vis.*, vol. 1, 1996, pp. 343–356.

[2] S. J. McKenna and H. Nait-Charif, "Tracking human motion using auxiliary particle filters and iterated likelihood weighting," *Image Vis. Comput.*, vol. 25, no. 6, pp. 852–862, Jun. 2007.

[3] C. J. Taylor, "Reconstruction of articulated objects from point correspondences in a single uncalibrated image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2000, pp. 677–684.

[4] G. Mori and J. Malik, "Recovering 3D human body configurations using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1052–1062, Jul. 2006.

[5] Y.-L. Chen and J. Chai, *3D Reconstruction of Human Motion and Skeleton From Uncalibrated Monocular Video*. Berlin, Germany: Springer, 2010.

[6] X. K. Wei and J. Chai, "Modeling 3D human poses from uncalibrated monocular images," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 1873–1880.

[7] X. Wei and J. Chai, "VideoMocap: Modeling physically realistic human motion from monocular video sequences," *ACM Trans. Graph.*, vol. 29, no. 4, p. 42, 2010.

[8] N. R. Howe, "Silhouette lookup for monocular 3D pose tracking," *Image Vis. Comput.*, vol. 25, no. 3, pp. 331–341, Mar. 2007.

[9] G. Shakhnarovich, P. Viola, and T. Darrell, "Fast pose estimation with parameter-sensitive hashing," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 750–757.

[10] R. W. Poppe, *Discriminative Vision-Based Recovery and Recognition of Human Motion*. Enschede, The Netherlands: Univ. Twente, 2009.

[11] M. Brand, "Shadow puppetry," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 2, Sep. 1999, pp. 1237–1244.

[12] N. R. Howe, M. E. Leventon, and W. T. Freeman, "Bayesian reconstruction of 3D human motion from single-camera video," in *Proc. Adv. Neural Inf. Process. Syst.*, 1999, pp. 820–826.

[13] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, Jun. 2001.

[14] A. Agarwal and B. Triggs, "3D human pose from silhouettes by relevance vector regression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun./Jul. 2004, p. 2.

[15] A. Agarwal and B. Triggs, "Recovering 3D human pose from monocular images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 44–58, Jan. 2006.

[16] A. Agarwal and B. Triggs, "Monocular human motion capture with a mixture of regressors," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Sep. 2005, p. 72.

[17] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, no. 1, pp. 79–87, 1991.

[18] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," in *Proc. IEEE/INNS Joint Conf. Neural Netw.*, Nagoya, Japan, 1993, pp. 181–214.

[19] C. Sminchisescu and A. Jepson, "Density propagation for continuous temporal chains. Generative and discriminative models," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, Tech., Rep. CSRG-401, 2004.

[20] C. Sminchisescu and A. Jepson, "Generative modeling for continuous non-linearly embedded visual inference," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, p. 96.

[21] R. Urtasun and T. Darrell, "Sparse probabilistic regression for activity-independent human pose inference," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[22] M. Isard and A. Blake, "CONDENSATION—Conditional density propagation for visual tracking," *Int. J. Comput. Vis.*, vol. 29, no. 1, pp. 5–28, 1998.

[23] T. Rui, Q. Zhang, Y. Zhou, and J. Xing, "Object tracking using particle filter in the wavelet subspace," *Neurocomputing*, vol. 119, pp. 125–130, Nov. 2013.

[24] X. Liu and S. Payandeh, "A study of chained stochastic tracking in RGB and depth sensing," *J. Control Sci. Eng.*, vol. 2018, no. 6, pp. 1–10, 2018.

[25] H. Sidenbladh, M. J. Black, and D. J. Fleet, "Stochastic tracking of 3D human figures using 2D image motion," in *Proc. Eur. Conf. Comput. Vis.*, 2000, pp. 702–718.

[26] A. Doucet, "On sequential simulation-based methods for Bayesian filtering," Dept. Eng., Univ. Cambridge, Cambridge, MA, USA, Tech. Rep. TR.310, 1998.

[27] A. Doucet, N. De Freitas, K. Murphy, and S. Russell, "Rao-blackwellised particle filtering for dynamic Bayesian networks," in *Proc. 6th Conf. Uncertainty Artif. Intell.*, Jun. 2000, pp. 176–183.

[28] G. Casella and C. P. Robert, "Rao-Blackwellisation of sampling schemes," *Biometrika*, vol. 83, no. 1, pp. 81–94, 1996.

[29] X. Xu and B. Li, "Exploiting motion correlations in 3-D articulated human motion tracking," *IEEE Trans. Image Process.*, vol. 18, no. 6, pp. 1292–1303, Jun. 2009.

[30] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Proc. Int. Stat. Optim. Perspect. Workshop 'Subspace, Latent Struct. Feature Selection'*, 2005, pp. 34–51.

[31] H. Sidenbladh, M. J. Black, and L. Sigal, *Implicit Probabilistic Models of Human Motion for Synthesis and Tracking*. Berlin, Germany: Springer, 2002.

[32] R. Urtasun, D. J. Fleet, A. Hertzmann, and P. Fua, "Priors for people tracking from small training sets," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, vol. 1, Oct. 2005, pp. 403–410.

[33] R. Urtasun, D. J. Fleet, and P. Fua, "3D people tracking with Gaussian process dynamical models," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 238–245.

[34] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 283–298, Feb. 2008.

[35] Z. Lu, C. Sminchisescu, and M. Á. Carreira-Perpiñán "People tracking with the Laplacian Eigenmaps latent variable model," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2007, pp. 1705–1712.

[36] R. Li, T.-P. Tian, S. Sclaroff, and M.-H. Yang, "3D human motion tracking with a coordinated mixture of factor analyzers," *Int. J. Comput. Vis.*, vol. 87, nos. 1–2, pp. 170–190, 2010.

[37] J. Liu, D. Liu, J. Dauwels, and H. S. Seah, "3D human motion tracking by exemplar-based conditional particle filter," *Signal Process.*, vol. 110, pp. 164–177, May 2015.

[38] I.-C. Chang and S.-Y. Lin, "3D human motion tracking based on a progressive particle filter," *Pattern Recognit.*, vol. 43, no. 10, pp. 3621–3635, Oct. 2010.

[39] S.-D. Bao, X.-L. Meng, W. Xiao, and Z.-Q. Zhang, "Fusion of inertial/magnetic sensor measurements and map information for pedestrian tracking," *Sensors*, vol. 17, no. 2, p. 340, 2017.

[40] M. Du, X. Nan, and L. Guan, "Monocular human motion tracking by using DE-MC particle filter," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3852–3865, Oct. 2013.

[41] L. Sigal, A. O. Balan, and M. J. Black, "HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *Int. J. Comput. Vis.*, vol. 87, nos. 1–2, pp. 4–27, 2010.

[42] M. P. Kumar, P. H. S. Torr, and A. Zisserman, "Extending pictorial structures for object recognition," in *Proc. Brit. Mach. Conf.*, BMVA Press, Sep. 2004, pp. 81.1–81.10.

[43] E. Poon and D. J. Fleet, "Hybrid Monte Carlo filtering: Edge-based people tracking," in *Proc. Workshop Motion Video Comput.*, Dec. 2002, pp. 151–158.

XIUKAI ZHAO received the bachelor's degree in computer science and technology from Shandong Normal University, Jinan, China, in 2016, where he is currently pursuing the master's degree with the School of Information Science and Engineering. His current research interests include virtual reality and artificial intelligence.

LEI LYU received the Ph.D. degree in computer application technology from the University of Chinese Academy of Sciences, in 2013. He is currently an Associate Professor with the School of Information Science and Engineering, Shandong Normal University, Jinan, China. His current research interests include virtual reality and computer vision.

JINLING ZHANG is currently an Associate Professor with the School of Information, Renmin University of China, Beijing, China. Her current research interests include virtual reality and multimedia.

CHEN LYU received the Ph.D. degree in computer application technology from the University of Chinese Academy of Sciences, in 2015. He is currently an Associate Professor with the School of Information Science and Engineering, Shandong Normal University, Jinan, China. His current research interests include artificial intelligence and natural language processing.