



Kinematics-based tracking of human walking in monocular video sequences

Huazhong Ning*, Tieniu Tan, Liang Wang, Weiming Hu

*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences,
Zhongguancun East Road 95, Beijing 100080, China*

Received 19 June 2003; received in revised form 5 January 2004; accepted 15 January 2004

Abstract

Human tracking is currently one of the most active research topics in computer vision. This paper proposed a kinematics-based approach to recovering motion parameters of people walking from monocular video sequences using robust image matching and hierarchical search. Tracking a human with unconstrained movements in monocular image sequences is extremely challenging. To reduce the search space, we design a hierarchical search strategy in a *divide-and-conquer* fashion according to the tree-like structure of the human body model. Then a kinematics-based algorithm is proposed to recursively refine the joint angles. To measure the matching error, we present a pose evaluation function combining both boundary and region information. We also address the issue of initialization by matching the first frame to six key poses acquired by clustering and the pose having minimal matching error is chosen as the initial pose. Experimental results in both indoor and outdoor scenes demonstrate that our approach performs well.

© 2004 Published by Elsevier B.V.

Keywords: Kinematics-based Tracking; Gait recognition; Human model

1. Introduction

Visual analysis of human motion is currently one of the most active research topics in computer vision. It concerns the detection, tracking and recognition of people, and more generally, the understanding of human behaviors, from image sequences involving humans. Human tracking, a hard but important task in human motion analysis, aims to recover continuously the global positions of the subject in the sequence, and more challengingly, to recover the joint angles in each frame. There have been many approaches to solving this problem, e.g. feature-based tracking [20,21], region-based tracking [22,23], active contour-based tracking [24], exemplar-based tracking [25,26], and model-based tracking [3–7,9,11]. The tracking results are widely applicable in many domains such as virtual reality, sports performance analysis and athlete training, the clinical study of orthopedic

patients, computer-driven rehabilitation environments, choreography, smart surveillance systems, gesture-driven user interfaces, video annotation, etc. [1,18,19].

Tracking human in video sequences is a very difficult task [12]. The difficulties can be categorized as internal issues and external issues. The internal issues are derived from the complex non-rigid structure of the human body. It has many joints and each body part can move in a wide range around its corresponding joint. So human motion involves a large number of degrees of freedom (DOFs) (about 34 for a full body) and frequent self-occlusions of body parts. The external issues result from the sensors, clothes, and background. Sensors of low quality will produce noisy images. Cluttered background and changing brightness often add inaccuracy to the motion segmentation, and the monotone clothes will worsen it. To address all of these issues in a realistic system using current vision technologies is impossible. So nearly all previous work, including this paper, makes some assumptions and copes with a simplified problem. Here, we assume that there is only one subject walking parallel to the image plane in each frame. This assumption is reasonable for the intended

* Corresponding author. Tel.: +86-10-626-47441; fax: +86-10-625-51993.

E-mail addresses: hzning@nlpr.ia.ac.cn (H. Ning), tnt@nlpr.ia.ac.cn (T. Tan), lwang@nlpr.ia.ac.cn (L. Wang), wmhu@nlpr.ia.ac.cn (W. Hu).

application of gait recognition because only one subject is often captured walking parallel to the image plane when the camera is installed in a desirable configuration for obtaining richer information of gait motion.

To address some internal issues, much previous work adopted human body models of various complexities to sufficiently use the prior knowledge relevant to the physical structure of the human body (see surveys [1,2,8,19] for more information). In earlier research, stick figure model was frequently used [3]. The simple stick figure model represents the human body parts by sticks that are connected by joints. More complex volumetric models, such as cylinder [4,9], truncated cone [5,11] and super quadrics [6], were used in later work. Recently, Plankers and Fua [7] presented a hierarchical human model including four levels: skeleton, ellipsoid metaballs simulating tissues and fats, polygonal surface representing skin, and shaded rendering. In general, the more complex the human body model is, the more precise the tracking results are. But a complex human body model leads to high computational cost. As a trade-off, we adopt an articulated truncated cone model with the head represented by a sphere in this paper.

Besides the human body model, image information used in PEF also varies. It includes boundary, region and texture, silhouette and contour, sticks and joints, blobs, depth, and so on. The most widely used image information is perhaps the boundary because it can be accurately localized and easily acquired [4,14]. Another one is the region which employs more information of the image and therefore achieves more robust results [15]. In this paper, we combine both boundary and region information in PEF to achieve both accuracy and robustness.

Corresponding to this specific PEF, we adopt a kinematics-based approach to pose refinement. In practice, a tracking procedure is often involved with two stages: pose prediction and pose refinement [1]. Prediction is usually realized through a specific dynamic model. As to the refinement stage, it should be carefully designed to reduce the solution space, since searching an optimal or approximately optimal pose in a high-dimensional body configuration space is intrinsically difficult. Generally, three main categories of refinement exist: kinematics, Taylor models and stochastic sampling. Kinematical approaches use physical forces applied to each rigid part of the body model of the tracked object. The forces as heuristic information guide the minimization of the difference between the pose of the body model and the pose of the real object [5,11]. Taylor models incrementally improve an existing estimation, using differentials of motion parameters with respect to the observation to predict better search directions [35]. It at least finds local minima, but cannot guarantee global optimality. Stochastic sampling handles the tracking in a probabilistic framework. It evaluates the posterior distribution by factored sampling and then propagates the sample set over time [27,29].

The three kinds of methods have their own advantages and disadvantages. Taylor models, as a gradient-based optimization, has fast convergence speed and low computational cost, but demands a differentiable PEF, which usually cannot be satisfied in such a complex problem as human tracking. Stochastic sampling maintains a sample set to represent simultaneous alternative hypotheses and propagates it over time so that it can cope with clutters in the images. But the dimension of the sample set increases exponentially as the scale of the problem increases, so the computational cost will be very high for human tracking. Kinematics-based method, like the Taylor model, cannot guarantee global optimality, but is superior to it with respect to the differentiability constraints. It also requires much lower computational cost than the stochastic sampling methods. Its advantages will be demonstrated in this paper.

2. Outline of our approach

Our kinematics-based tracking of human walking is illustrated in Fig. 1. To improve the speed of minimization of the PEF, we design a hierarchical search strategy since the articulated human body model is naturally formulated as a tree-like structure. This strategy includes two stages: location of global position and estimation of joint angles. We search the global position in each frame by finding the centroid of the human body. To estimate the joint angles, after prediction using a dynamic model, physical forces similar to those in Refs. [5,11] are used as heuristic information to recursively refine the predicted body poses. Since in the first frame there is no tracking history information for prediction, a specific initialization process is run instead. For the tracking of walking human, our initialization requires lower computational cost or produces better results compared with much previous work [9,12,13,30] (see Section 5.3). This kinematics-based approach can be explained intuitively and has the advantage of low computational cost.

The tracking results including position, orientation and joint angles of each frame in video sequences are intended to be used for gait recognition. There are many properties of gait that might serve as recognition features. We can categorize them as static features and dynamic features. The former usually reflects geometry-based measurements such as body height and build, while the latter means lower-body joint-angle trajectories. Intuitively, recognizing people by gait depends greatly on how the static silhouette shape of an individual changes over time. So previous work on gait recognition mainly adopted low-level information such as silhouette [16,17] and few methods used higher-level information such as temporal features of joint angles. Because they are more essential to sufficiently reflect the dynamics of gait motion, we expect to try these dynamic gait features for personal recognition, or use fusion of both dynamic and static features available from body biometrics.

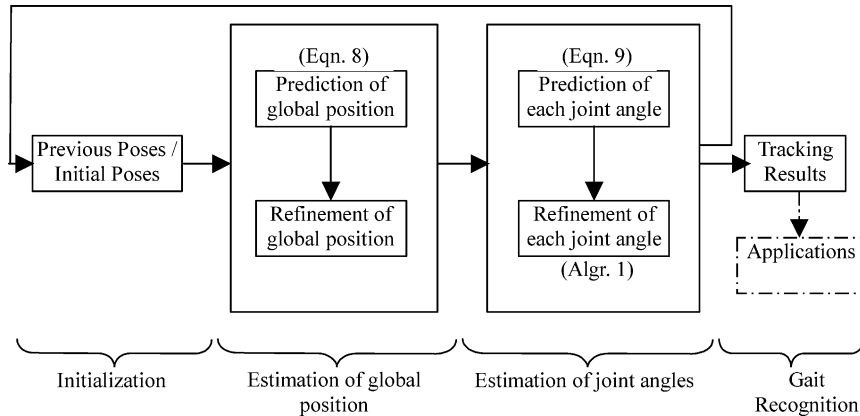


Fig. 1. Framework of our approach.

Though our final goal is to recognize people by their gaits using vision techniques, this paper is focusing on dynamic data acquisition by tracking walkers in video sequences.

This paper is an extended version of our previous work described in Ref. [33]. The major modification lies in human detection, boundary matching error, refinement of joint angles, initialization, and gait recognition. The main contributions of this paper are summarized as follows:

- A pose evaluation function (PEF) combining both boundary and region information is proposed that derives the accuracy from the former and robustness from the latter.
- To reduce the search space and according to the tree-like structure of human body model, we design a hierarchical search strategy in a *divide-and-conquer* fashion to decompose the parameter space so that the human global position and all joint angles can be estimated separately.
- A kinematics-based algorithm is proposed to recursively refine the joint angles, which is easier to implement than the gradient-based approach and has a lower computational cost than the popular CONDENSATION method in human tracking.
- We address the issue of initialization in an effective way. In initialization, the first frame is matched to six key poses acquired by clustering and the pose having minimal matching error is chosen as initial pose.
- Dynamic features of individual gait (i.e. joint-angle trajectories) are used to identify people and the recognition rate seems better than our other work [31] using static features.

The remainder of this paper is arranged as follows. Section 3 details the human body model. Section 4 presents a PEF combining both boundary and region information. Section 5 designs a hierarchical search strategy to decompose the parameter space. Section 6 gives some experimental results and discussions. The paper is concluded in Section 7.

3. Human body model

Our human body model, similar to Refs. [5,9,11], is composed of 14 rigid body parts, including upper torso, lower torso, neck, two upper arms, two lower arms, two thighs, two legs, two feet and a head. Each body part is represented by a truncated cone except for the head represented by a sphere. They are connected to others at joints, whose angles are represented as Euler angles. We do not model hands because they are very complex and are of little importance in human body tracking. Fig. 2 gives some perspective views of the human body model used in this paper. This is a generic model. But for person-specific tracking, we must adjust its dimensions to individualize the model.

Without considering the static parameters (including the shape and size of each part), the above human body model in its general form still has 34 DOFs of dynamic parameters: 2 DOFs for each body part (14×2), 3 DOFs for its global position (translation), and 3 DOFs for its orientation (rotation). To search quickly in a 34-dimensional state space is extremely difficult. However, in the case of gait recognition, people are usually captured walking parallel to the image plane when the camera is installed in a desirable configuration, and the movements of the head, neck and lower torso relative to the upper torso are very small. Therefore, the state space can be naturally reduced with such constraints. In this paper, we assume that only the arms and legs have relative movements while the upper torso

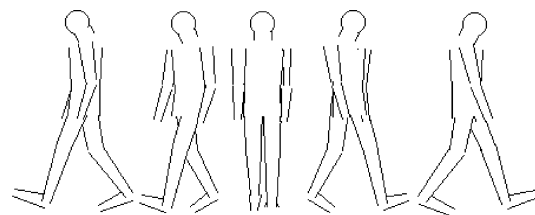


Fig. 2. Human body model projected into the image plane from five viewing angles.

moves approximately along a line; and each joint has only one DOF. This reduces the dimensionality of the state space to 12: 1 DOF for each joint mentioned above plus 2 DOFs for the global position. We represent the global position and joint angles by a 12-dimensional state vector $P = \{x, y, \theta_1, \theta_2, \dots, \theta_{10}\}$ where (x, y) is the global position of the human body and θ_i is the i th joint angle. This state vector describes the relative position of different body parts; and the intention of tracking is to recover the state $P = \{x, y, \theta_1, \theta_2, \dots, \theta_{10}\}$ from each frame, and then the recovered parameters are used as dynamic features for gait recognition.

In model-based tracking, we need to synthesize and project the body model onto the image plane given the camera parameters and state vector $P = \{x, y, \theta_1, \theta_2, \dots, \theta_{10}\}$. In other words, we need to calculate the model coordinates of each point on the body model, and transform them to the camera coordinates and then to the image coordinates. To locate the positions of model parts in the model coordinate system, each part is defined in a local coordinate frame with the origin at the base of the truncated cone (or center of the sphere). Each origin corresponds to the center of rotation (i.e. the joint position). We represent the human body model as a kinematical tree (with the torso at its root) to order the transformation between the local coordinate frames of different parts. Therefore, the model coordinates of each part are formulated as the product of transformation matrices of all the local coordinates on the path from the root of the kinematical tree to that part.

In detail, for each body part s_i on the kinematical tree, there is a path $s_i s_{i-1} \dots s_0$ starting from s_i and ending at the root s_0 . For each point p on s_i , we can easily calculate its position X_i in the local coordinates. Then after translation and rotation, X_i is transformed to X_{i-1} in the local coordinates of body part s_{i-1} , furthermore to X_{i-2} in that of s_{i-2}, \dots , and finally to X_0 in that of s_0 . X_0 are also the model coordinates of that point p on s_i . So the key is to calculate the transformation matrix between two connected body parts. In Fig. 3, part 2 has a translation $t = (t_x, t_y, t_z)$ and a rotation $(\theta_x, \theta_y, \theta_z)$ from part 1. For each point on part 2, given its position $X = (x, y, z, 1)$ in the local coordinates of part 2, the transformation of X to X' in the local coordinates of part 1 is

$$X' = \mathbf{R}_x(\theta_x)\mathbf{R}_y(\theta_y)\mathbf{R}_z(\theta_z)\mathbf{T}(t)X \tag{1}$$

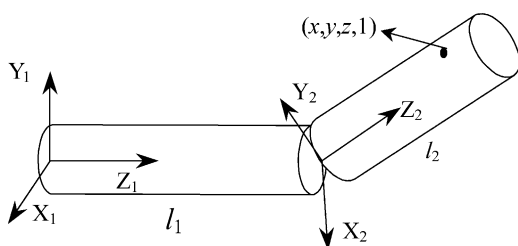


Fig. 3. Coordinate transformation between two connected body parts.

where $\mathbf{R}_x, \mathbf{R}_y, \mathbf{R}_z$ are rotation matrices and \mathbf{T} is translation matrix. We assume that the world coordinate system is superposed on the model coordinate system and the camera is modeled as a pinhole camera. Given the projection matrix \mathbf{M} , each point on the body model with model coordinates $X = (x, y, z, 1)$, after projection, has the image coordinates $X_i = \mathbf{M}X/z$.

4. Pose evaluation function

The main task of our model-based tracking is to relate the image data to the pose vector $P = \{x, y, \theta_1, \theta_2, \dots, \theta_{10}\}$. The general method to this problem is known as *analysis-by-synthesis*, and is used in *predict-match-update* fashion [1]. The philosophy, in detail, is to predict the most possible pose of the human body model in the current frame according to the tracking results in the previous frames. Then the human model with the predicted pose is projected into the image plane. We match the model projection to the edge image extracted by human motion detection (see Section 6.2) and measure the matching error by a PEF. Next the predicted pose is recursively refined to minimize the PEF. So PEF plays an important role in model-based tracking. Our PEF is a combination of boundary matching error and region matching error. As the PEF is computed, physical forces are easily generated and used to recursively refine the estimation of joint angles (see Section 5).

4.1. Boundary matching

Our boundary matching error can be regarded as Chamfer distance [28] between the edges of model projection and edges extracted by human detection. It is also similar to that in Ref. [29]. Given the human body model with the pose $P = \{x, y, \theta_1, \theta_2, \dots, \theta_{10}\}$, we suppose that the boundary curve of model projection is $r(s), 0 \leq s \leq 1$ and the corresponding curve in the image data is $z(s)$. The mapping function $g(s)$ associates each point $z(s)$ on the curve in the image data with the corresponding point $r(g(s))$ on the boundary curve of model projection. In practice, $g(s)$ is not necessarily injective because $z(s)$ includes clutter as well as foreground features; and the distance between point $z(s)$ and $r(g(s))$ is limited to no greater than a spatial scale constant μ . So the boundary matching error E_b is defined as

$$E_b = \frac{1}{c} \int_0^1 \min(\|z_1(s) - r(s)\|, \mu) ds \tag{2}$$

where c is a normalization constant usually set to the length of the curve $r(s), 0 \leq s \leq 1$ and $z_1(s)$ is the closest associated feature to $r(s)$:

$$z_1(s) = z(s'), \quad s' = \operatorname{argmin}_{s' \in g^{-1}(s)} \|r(s) - z(s')\| \tag{3}$$

Fig. 4 shows the procedure of computing Eq. (2). For each pixel p_i on the boundary curve $r(s)$ of the model projection,

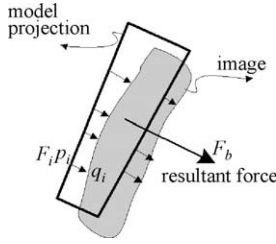


Fig. 4. Measuring the boundary matching error and its physical forces.

we search the corresponding pixel in the edge image along the gradient direction at pixel p_i . In other words, the pixel nearest to p_i and along the normal direction is the closest associated feature to p_i . Assuming that q_i is the corresponding pixel and that \mathbf{F}_i stands for the vector $\overrightarrow{p_i q_i}$, we regard the norm $\|\mathbf{F}_i\|$ as the matching error of pixel p_i to q_i . It is supposed that the closest associated feature to p_i is missing due to noises, if q_i does not exist or $\|\mathbf{F}_i\|$ is too big. In this case, the matching error of p_i is set to the constant μ . Finally, the boundary matching error E_b is the average of the matching errors of all pixels on the boundary of the model projection.

To provide the refinement of joint angles with heuristic information, we adopt the idea of spring forces [5,11]. The physical forces can be easily calculated here and will be used in Section 5. According to this idea, each \mathbf{F}_i described above is viewed as a spring with its end points attached to p_i and q_i , and each spring gives a physical force in proportion to $\|\mathbf{F}_i\|$, pulling p_i to q_i . Then the combination of all the physical forces, i.e. \mathbf{F}_b in Fig. 4, pulls the model projection to the corresponding image data.

$$\mathbf{F}_b = \frac{1}{c} \int_0^1 f(\mathbf{F}(s), \rho \frac{\mathbf{F}(s)}{\|\mathbf{F}(s)\|}) ds \quad (4)$$

where $\mathbf{F}(s) = \overline{r(s)z_1(s)}$, ρ is a spatial scale constant, and

$$f(\mathbf{F}_1, \mathbf{F}_2) = \begin{cases} \mathbf{F}_1 & \|\mathbf{F}_1\| \leq \|\mathbf{F}_2\| \\ \mathbf{F}_2 & \|\mathbf{F}_1\| > \|\mathbf{F}_2\| \end{cases}$$

4.2. Region matching

In general, the PEF defined in Eq. (2) can properly measure the similarity between the model projection and image data, but it is insufficient under certain circumstances. A typical example is given in Fig. 5(a), where the model projection falls into the gap between two body parts in the edge image. Although it is obviously badly fitted, a small matching error according to Eq. (2) may be obtained. To avoid such ambiguities, region information is further considered in our approach. Fig. 5(b) illustrates the region matching process. Here the region of the model projection, which is fitted into the image data, is divided into two parts: P_1 is the region overlapping with the image data and P_2

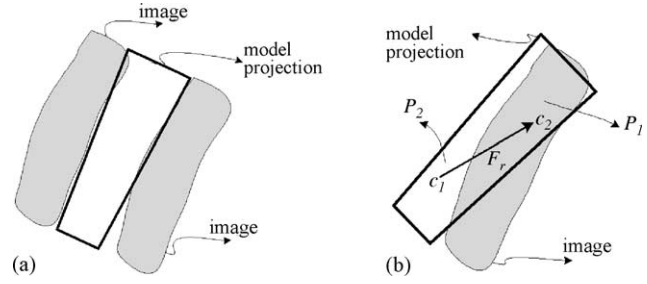


Fig. 5. (a) A typical ambiguity: a model part falls into the gap between two body parts in the image. (b) Measuring region matching error and its physical force.

stands for the rest. Then the matching error with respect to the region information is defined by

$$E_r = |P_2| / (|P_1| + |P_2|) \quad (5)$$

where $|P_i|$ is the area, i.e. the number of pixels in the corresponding region.

Similarly, another physical force is also defined. Supposing c_1 and c_2 are the centroids of the regions P_1 and P_2 , respectively, we define the vector $\mathbf{F}_r = \overrightarrow{c_1 c_2}$ as the physical force resulting from region matching. This physical force pulls the model projection to overlap the corresponding image part as greatly as possible.

In general, boundary information improves the localization, whereas region information stabilizes the tracking because more image information is used. Here we combine the two matching errors in the PEF $S(P)$ so as to achieve both accuracy and robustness; and so do the physical forces \mathbf{F}_b and \mathbf{F}_r . A factor α is used to adjust their weights:

$$S(P) = (1 - \alpha)E_b + \alpha E_r \quad (6)$$

$$\mathbf{F} = (1 - \alpha)\mathbf{F}_b + \alpha \mathbf{F}_r \quad (7)$$

where $P = \{x, y, \theta_1, \theta_2, \dots, \theta_{10}\}$ is the pose vector. How to determine the factor α is challenging. In general a smaller α is preferred for the upper limbs to alleviate the influence of region matching error. The reason is that the upper limbs and the torso often have clothes with the same texture and they frequently occlude each other, and therefore the region information is of relatively little importance. In our system, α is empirically selected as 0.6 for upper limbs and 0.8 for lower limbs to improve the weight of boundary matching. Fig. 6 shows a curve of our PEF. The surface of the evaluation function is basically smooth and has no local minima at the neighborhood of the global minimum. These two properties are desirable for optimization. Furthermore, according to the surface of the PEF, we find that the global position (x, y) is much more significant than other joint angles with respect to the matching error. This also explains why the global position is first predicted without considering other parameters in the following hierarchical search strategy.

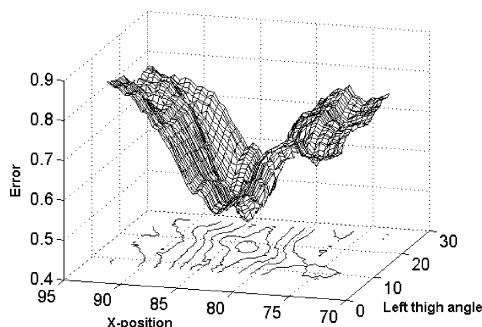


Fig. 6. The surface of the PEF with the global position x and the joint angle of the left thigh changing smoothly and other parameters remaining constant; also shown is the contour of the function.

5. Hierarchical search strategy

Our task of tracking is to search an optimal or suboptimal pose $P = \{x, y, \theta_1, \theta_2, \dots, \theta_{10}\}$ in the 12-dimensional body configuration space so as to minimize the PEF. But locating a pose in a high-dimensional space (e.g. 12 DOFs in our problem) is intrinsically difficult. Since the articulated human body model is naturally formulated as a tree-like structure, a hierarchical search strategy, i.e. locating the global position and tracking each limb separately, is suitable here. In detail, the global position (x, y) (the root of the tree) is firstly determined, and then we adopt a kinematics-based approach to estimate all joint angles on each path from the root to the leaf. As a divide-and-conquer method, our hierarchical search strategy decomposing the parameter space is supported by the following reasons. Firstly, the global position (x, y) is much more significant than other joint angles with respect to the PEF, so it can be estimated separately with other motion parameters fixed. Secondly, joint parameters are greatly dependent on the global position (x, y) . A slight deviation of the global position often causes the joint parameters to drastically deviate from their real values when minimizing the PEF. So the global position should be located before the estimation of the joint angles. Finally, because usually the upper limbs can hardly be segmented from the torso in the image, tracking the upper limbs is more difficult than tracking the lower limbs and they need to be considered separately.

5.1. Estimation of global position

Estimation of the global position includes two stages: prediction and refinement. In the prediction stage, we assume that change of the centroid of the human body between consecutive frames is equal to the change of its global position

$$\mathbf{X}_c - \mathbf{X}_p = \mathbf{C}_c - \mathbf{C}_p \text{ or } \mathbf{X}_c = \mathbf{C}_c - \mathbf{C}_p + \mathbf{X}_p \quad (8)$$

where the subscripts c and p indicate the current frame and the previous frame, respectively, and \mathbf{X} and \mathbf{C} mean the global position and the centroid, respectively. So, predicting

the global position of moving human in the current frame can be viewed as the problem of approximately calculating the centroid. According to the experiments on the SOTON and NLPR gait databases, the prediction error is mostly less than 3 pixels. So in the refinement stage, we completely search the neighborhood of the predicted value to minimize the PEF. If the prediction error is greater, a more powerful method such as meanshift [36] can be applied to the refinement stage. In Fig. 7, the global position is estimated at $(x, y) = (48, 12)$ (the origin of the image coordinate system is located at the center of the image), and the human body model with the joint angles equal to those in the previous frame is projected onto the image plane at this position. It can be seen that the estimation of global position is fairly accurate although the arms and legs deviate from the real position and the joint angles need to be further refined, which is the focus of Section 5.2.

5.2. Estimation of joint angles

Estimation of joint angles also has two stages: prediction and refinement. To roughly predict the joint angles, we need to calculate their angular speeds according to the tracking results in the previous frames and apply them to the kinematical equation

$$\theta_{ic} = \theta_{ip} + \dot{\theta}_{ip} \Delta t \quad (9)$$

where θ_i is the i th angle joint, the subscripts c and p indicate the current frame and previous frame, respectively, and $\dot{\theta}_{ip}$ stands for the angular speed in the previous frame. Then θ_{ic} is the predicted value of that joint angle. Here, we suppose that the angular speed is constant in a short time interval between two consecutive frames. Eq. (9) has a linear format but the dynamic model is essentially nonlinear, because the angular speed is not a constant but dynamically updated according to the tracking results in the previous frames. In practice, θ_{ic} and $\dot{\theta}_{ip}$ can basically reveal the dynamics of the gait oscillation. After prediction by the dynamic model, θ_{ic} is then recursively refined using the following kinematics-based approach.

In the hierarchical search strategy, the state of each son-node on the kinematical tree is estimated after the state of its



Fig. 7. After the global position is estimated at $(x, y) = (48, 12)$, the human body model with the joint angles equal to those in the previous frame is projected onto the image plane.

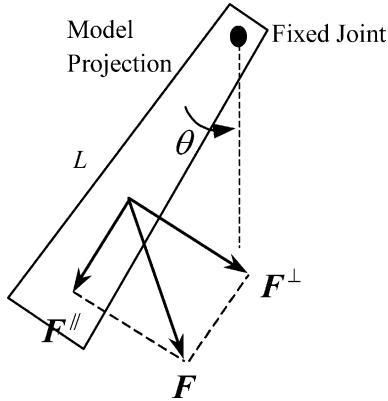


Fig. 8. Model projection pulled by the physical forces is rotating around the fixed joint.

parent is determined. So the position of the joints is fixed before we estimate their rotation angles. In this case, the kinematics-based approach is suitable to recursively refine the joint angles. In detail, the projection of the model part, with one end fixed at the joint, is pulled by the physical forces to rotate around the joint and to approximate the corresponding part in 2D image until the PEF reaches a minimum. In Fig. 8, the model projection with the length L is pulled by \mathbf{F} described in Eq. (7) that acts on the centroid of the model. According to the law of rotational motion of a rigid body about a fixed axis, the torque \mathbf{M} produced on the rigid body equals the rotational inertia I of the rigid body about the axis multiplied by the angular acceleration $\ddot{\theta}$.

$$I\ddot{\theta} = \mathbf{M} = \mathbf{r} \times \mathbf{F} \quad (10)$$

where \mathbf{r} is the arm of force. It is assumed that $r \propto L$ since the model projection approximates a rectangle and \mathbf{F} acts on its centroid. The rotational inertia $I = mk^2$ where m is mass and k is the radius of gyration. With the rectangle approximation, we also have $k \propto L$, and then $I \propto L^2$. Given that F^\perp is a component of the physical force \mathbf{F} vertical to the model projection, Eq. (10) has a scalar form

$$\ddot{\theta} \propto F^\perp / L \quad (11)$$

Pulled by the physical force \mathbf{F} , the model projection is rotating from stillness. After a short time interval Δt , the rotation angle $\Delta\theta = \frac{1}{2}\ddot{\theta}\Delta t^2 \propto \frac{1}{2}F^\perp\Delta t^2/L$, or with Δt set to time unit,

$$\Delta\theta = \beta F^\perp / L \quad (12)$$

where β is a constant independent of F and L .

In the refinement stage, we recursively update the joint angle of each model part by adding the increment $\Delta\theta$ in Eq. (12) to the last updated value. Then, a question arises: will the recursion always converge to the real value of the joint angle? In experiments, we found that sometimes the refined value would oscillate around the real value because a big β introduces a big increment. It is true that a smaller β

will avoid the oscillation to some extent. But a small β will greatly increase the number of iterations. So we adopt an adaptive recursion by reducing β as the number of iterations k increases, i.e. $\beta = \gamma/k$, where γ is a constant scalar. The recursion terminates when the matching error is less than a threshold δ , the number of iterations reaches K or the amplitude of the increment $\Delta\theta$ is less than a threshold ε . The major steps of estimation of the joint angles are summarized as follows.

Algorithm 1.

1. Predict the joint angle θ^0 according to Eq. (9) and set $k = 0$;
2. Compute the physical force \mathbf{F} according to Eq. (7).
3. Calculate the increment $\Delta\theta$ according to Eq. (12). If $|\Delta\theta| < \varepsilon$, the recursion is stopped, otherwise goes to step 4.
4. $\theta^{k+1} = \theta^k + [\beta/(k+1)]\Delta\theta$ and the joint angle is updated to θ^{k+1} .
5. Project the human body model with the updated pose P onto the image plane and compute the PEF $S(P)$ according to Eq. (6).
6. If $S(P) < \delta$ or $k \geq K$, the recursion is stopped, otherwise goes to step 2.

5.3. Initialization

In the above hierarchical search strategy, the poses in the previous frames are needed to predict the current pose. But as to the first frame, the previous tracking results are unavailable and a specific initialization process roughly estimating the human pose is thus necessary to replace the prediction stage in the above algorithm 1. Many previous approaches handled initialization by manually adjusting the human body model to approximate the real pose or by arbitrarily assuming that the initial pose is subject to a uniform distribution [9,13,30]. Cheng and Moura [12] provided an automatic initialization method that searched the entire motion model to locate the first frame by finding the dominant peak of a cost function. However, this approach evaluates the cost function many times, leading to high computational cost. Our approach to initialization is similar to Cheng's work, but with lower computational cost.

We cluster the training data of human poses acquired manually. In each walking cycle, six clusters exist and the mean of each cluster is the key pose (see Fig. 9). To initialize the first frame, we simply evaluate the six key poses using Eq. (6) and the one having minimal value is chosen as the initial pose. Fig. 10 is an example of initializing the first frame. It can be seen that the estimated initial pose (cluster 2) is fairly accurate except some small errors in detail. Then the initial pose is refined using the algorithm in Section 5.2. It can be found that the PEF needs

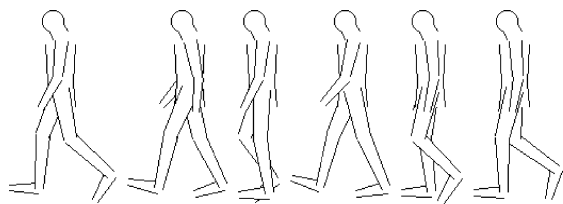


Fig. 9. Clusters of human poses.

to be evaluated only six times, which naturally results in much lower computational cost compared with Cheng's work. This initialization method can also be used to recover from severe tracking failures due to occlusion, accumulated error, or image noise. When a severe failure occurs (when the matching error reaches a predefined threshold), the tracker will stop temporarily and reinitialize the current frame.

6. Experimental results

To verify the effectiveness of our approach, we have carried out a large number of experiments on video sequences in both indoor and outdoor scenes. The experimental results and detailed discussions are described as follows.

6.1. Data acquisition

The experiments are carried out on image sequences captured in both indoor and outdoor environments. For the indoor scene, we use the earlier SOTON gait database [18]. The database includes six subjects and four sequences of each subject. These sequences were captured by a fixed camera with a stationary indoor background, at a rate of 25 frames per second, and the original resolution is 384×288 pixels. The outdoor sequences are captured at the same frame rate by a digital camera (Panasonic Nv-Dx100EN) fixed on a tripod and the original resolution of these images is 352×240 pixels. These outdoor sequences form the NLPR gait database that includes

20 subjects and four sequences per subject. Some samples are showed in Fig. 11.

6.2. Moving human detection

Moving human detection is the first step of our approach to track a walking person. To extract walking figures from the background image, background subtraction is adopted that is a particularly efficient method for detecting changes where a fixed camera is usually used to observe dynamic scenes. Generally, motion detection based on background subtraction involves background modeling, the arithmetic subtraction operation and the selection of a suitable threshold.

In this paper, the least median of squares (LMedS) method is used to reliably model the background image [10]. Its advantages are that it is efficient especially for 1D data such as a pixel process in returning the correct result even when a large amount of outliers are present. The brightness change is usually accomplished by differencing between the background and current image. However, the selection of threshold for binarization is very difficult, especially in the case of low contrast images as most of moving objects may be missed out since the brightness change is too low to distinguish changing regions from noise. To solve this problem, we choose an extraction functions to indirectly perform differencing operation for gray images and color images respectively (see Ref. [31] for detailed information). After binarization, we eliminate the outliers using morphological operators such as dilating, eroding and hole-filling to smooth the result. Then by masking the original image with the binary image, the region of moving human is accurately obtained (see Fig. 12(c)). Finally, using the Sobel operator, we acquire the edges of the moving human (see Fig. 12(d)), which is used in matching.

6.3. Tracking results

After the first frame in each sequence is initialized, the tracker uses the proposed hierarchical search strategy to

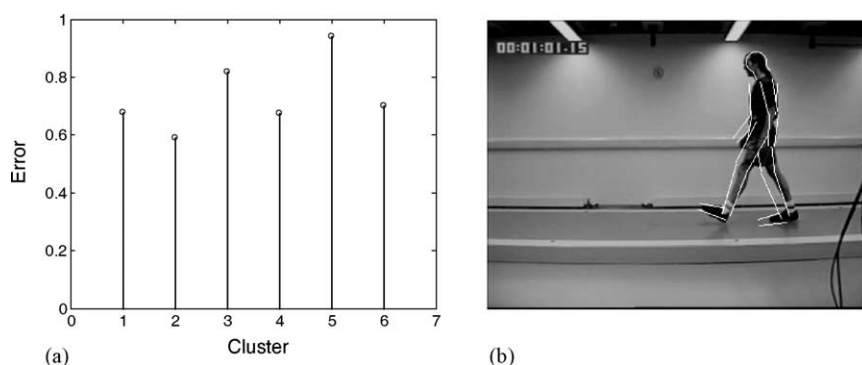


Fig. 10. An example of initializing the first frame. (a) Matching errors of the 6 clusters computed by Eq. (6) and cluster 2 having minimal error. (b) The human body model with the joint angles equal to cluster 2 is projected onto the image plane.

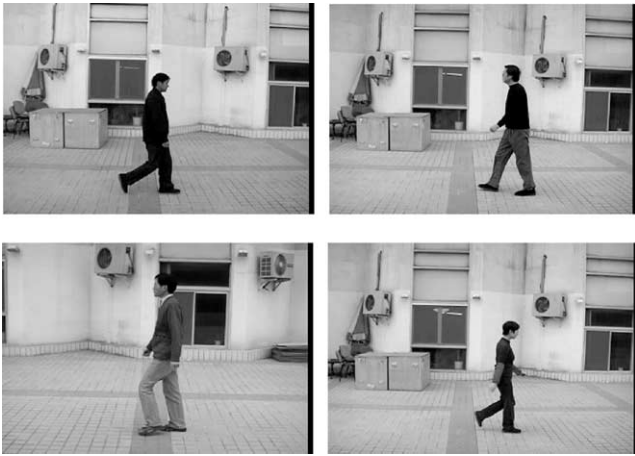


Fig. 11. Samples in NLPR gait database.

automatically recover the human pose $P = \{x, y, \theta_1, \theta_2, \dots, \theta_{10}\}$ from each image in the sequence. Here, we show two sequences of the tracking results (see Figs. 13 and 14). Due to the space constraint, only the human areas clipped from the original image sequences are shown. Some sequences include challenging configurations in which the two legs and thighs occlude each other severely (e.g. frame 11 in Fig. 13 and frame 25 in Fig. 14), causing most part of one leg or thigh is unseen. These challenging data verify the effectiveness of our approach. Other challenges include shadow under the feet, the arm and the torso having the same color, various colors and styles of clothes, different shapes of the tracked people, and low quality of the image sequences. It is noted that the arm far from the camera in both sequences was lost for the severe occlusion by the torso.

As mentioned above, our dynamic model in Eq. (9), used to predict the human pose in each frame, has a linear

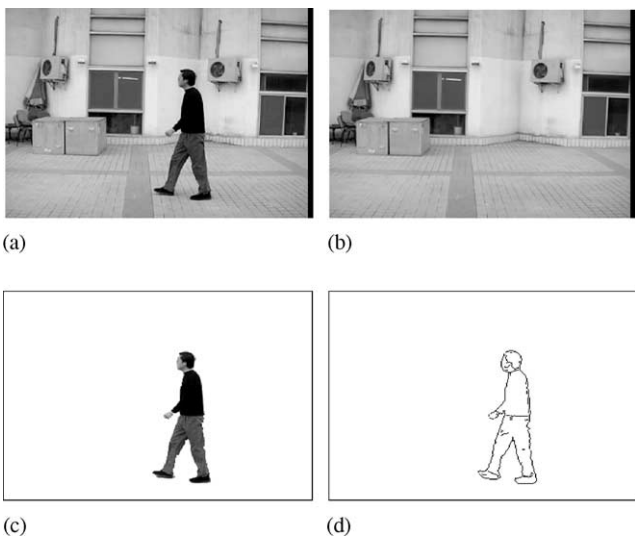


Fig. 12. Moving human detection. (a) Original image. (b) Background. (c) Image masked by binary image. (d) Edges of moving human.

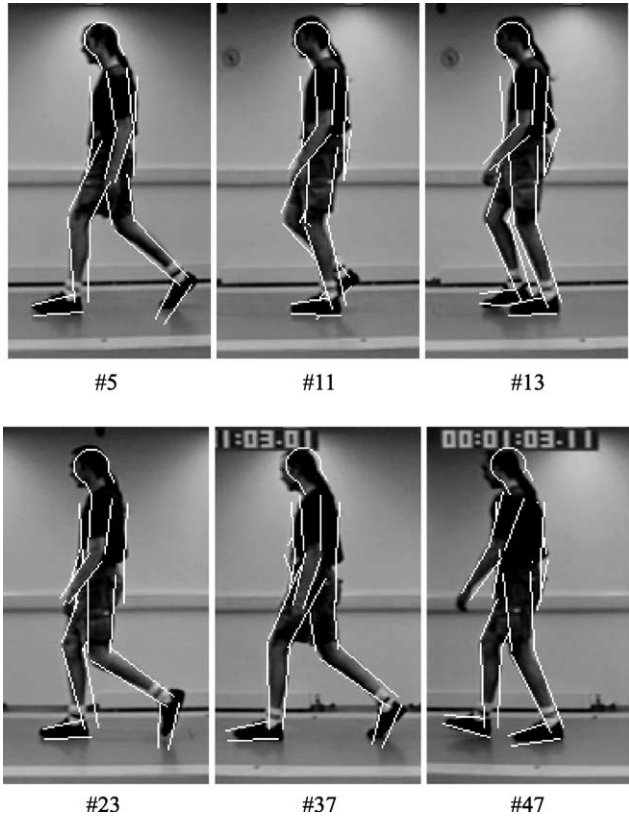


Fig. 13. Tracking results of indoor walking.

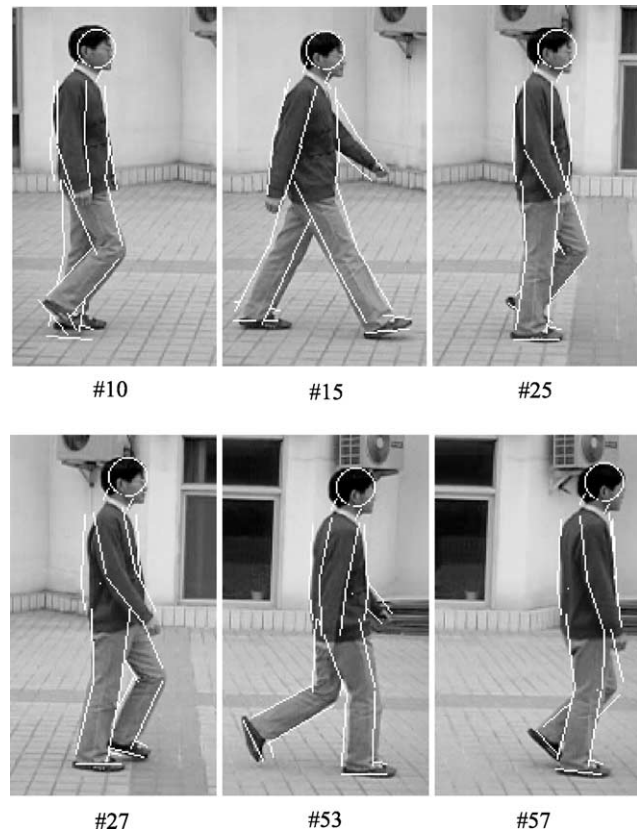


Fig. 14. Tracking results of outdoor walking.

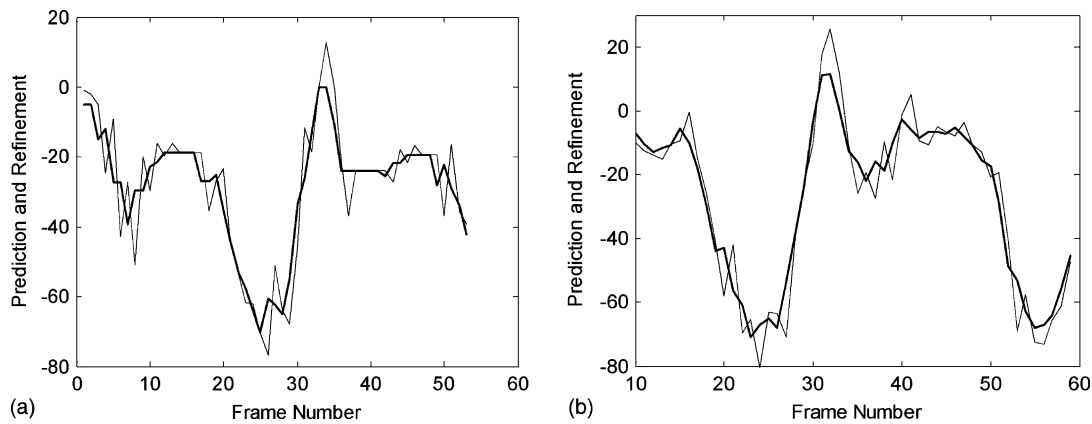


Fig. 15. Curves of prediction and refinement. (a) Curves of the joint of left leg corresponding to the sequence in Fig. 13. (b) Curves of the joint of left leg corresponding to the sequence in Fig. 14. In (a) and (b), the curve with thin line is the prediction curve and curve with thick line is the refinement curve.

format but nonlinear essence. It may be questioned that what role it plays in our tracking approach. To answer it, we show some temporal curves of the predicted and refined joint angles in Fig. 15. It can be seen that the prediction curves dithers drastically (represented by thin lines) but roughly fitting the refinement curve. After refined by the kinematics-based algorithm, the temporal curves of the joint angles are smoothed (represented by thick lines). So we can conclude that our dynamic model can roughly predict the human pose and the kinematics-based approach can effectively reduce the prediction errors.

6.4. Gait recognition

As mentioned earlier, the intended purpose of model-based tracking of human walking is to acquire motion parameters of walking (e.g. joint angles and velocity) for gait recognition. Gait, the manner of walking, is a newly emergent biometric which offers the possibility to recognize people at a distance without any interaction from the subject. Accordingly, vision-based automatic gait recognition is so attractive as a method of human identification from a surveillance perspective. The growing interest in automatic gait recognition has led to a significant progress over the past few years. For instance, Lee and Grimson [37] divided the silhouette region into seven sub-regions that did not meant the accurate segmentation of arms or legs and then made use of the moment features of image regions to recognize individuals; Yam et al. [38] used the temporal template matching to extract the rotation angles of the thigh and lower leg for gait recognition; and Tanawongsuwan and Bobick [39] attached markers to joints to acquire motion trajectories and then used the dynamic features derived from the trajectories of lower-body joint angles such as the hip and the knee to recognize individuals. It is true that these methods utilized dynamic features of human walking, but the former two methods could not obtain motion data of

high accuracy and the latter needs markers that often cannot meet the real application of gait recognition. So, in this paper we apply the model-based tracking to full-automatically acquire dynamic features of human gait for recognition.

Fig. 16 shows the temporal curves of thigh and knee angles corresponding to the sequences in Figs. 13 and 14 that are chosen as the dynamic features of individual gait for recognition. We have done some initial work on gait recognition on the SOTON and NLPR gait database using the dynamic features from the lower limbs. The Euclidean distance is used to measure the similarity of the features. For a small number of examples, we compute an unbiased estimate of the true recognition rate using a leave-one-out cross-validation method. That is, we leave one example out, train on the rest, and then classify or verify the omitted element according to its similarities with respect to the rest examples.

To measure the performance of gait recognition, we use the cumulative match characteristic (CMC) curve proposed in the face recognition community that indicates the probability that the correct match is included in the top n matches. For completeness, we also use the ROC curves to report verification results. Fig. 17 shows the performance of identification and verification using dynamic and static information, respectively. The recognition results using static features are detailed in our work [31]. It can be seen that the correct recognition rate and equal error rate (EER) using dynamic features are 87.5 and 8% that are better than the results (84 and 10%, respectively) using static features. So the dynamic features seem to contain richer information than the static features for gait recognition. In our most recent work [34], we tried the fusion of both dynamic and static features available from body biometrics for personal recognition and the achieved results are better than any single feature. The application of gait recognition also proves the usefulness of the proposed tracking algorithm.

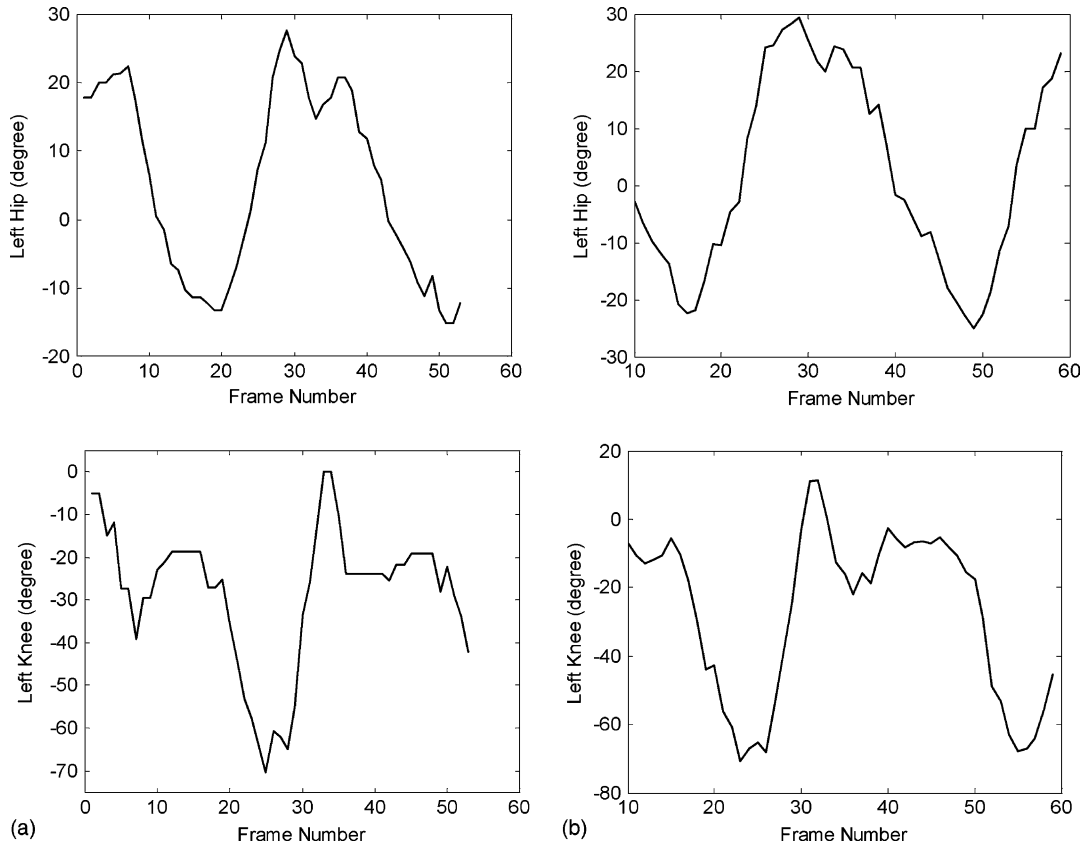


Fig. 16. Temporal curve of joint angles. Top row: left thigh angles; bottom row: left knee angles. (a) and (b) correspond to the sequences in Figs. 13 and 14, respectively.

6.5. Discussions and future work

Our kinematics-based approach has some advantages over the gradient-based search strategy [9,32]. Gradient-based optimization uses the gradient of the PEF to search the maximum or minimum, so the differentiability of the PEF is necessary. However, in complex problems like human body tracking in this paper, the PEF is rarely differentiable and various assumptions must be made to

approximate the gradient. As the errors accumulate, the approximation may be offset by the noise in images and the gradient will fail to direct the search. In contrast, the increment $\Delta\theta$ in the kinematics-based approach can be easily computed from the image and has obvious physical meaning to pull the model projection to match the image data, although it is at the expense of a little higher computational cost. While compared with CONDENSATION [27,29], the kinematics-based approach requires

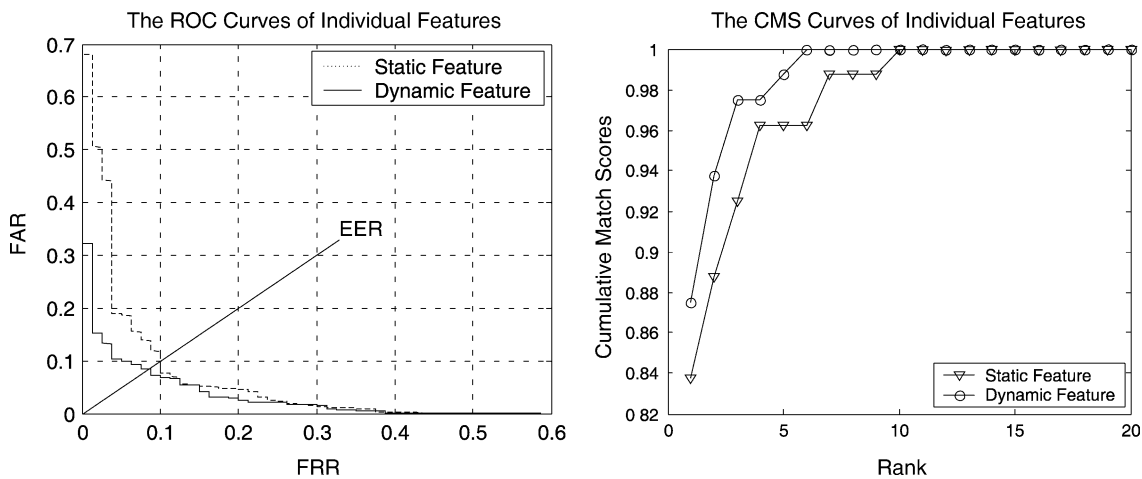


Fig. 17. Identification and verification performance.

much lower cost because a large sample set is needed for the former. Although the advantages, the proposed approach needs to be improved in the following aspects.

Walking direction: though sufficient for gait recognition, in which human usually walks parallel to the image plane by proper camera configuration, our approach is not appropriate for unconstrained movements. The chief reason is that the calculation and application of physical forces is based on the assumption that human walks parallel to the image plane. To remove this constraint, more powerful methods are needed, such as that in Refs. [5,11].

Motion model. In the tracking results, the arms far from the camera were completely lost. But the movements of two arms are generally symmetric in walking. So the pose of the occluded arm might be estimated according to the other arm using a motion model of walking. Also as prior knowledge, the motion model is powerful for prediction, representation and recognition of movements, especially when applied to periodical movements such as walking. In Ref. [12], temporal walking curve was represented by a 3-order B spline. A finite state machine was learnt to track the structured motion of ballet in Ref. [13]. Their effective results encourage us to focus on motion model in future work.

Texture model. Although region information was considered in our approach to measure matching error and to remove the ambiguities described in Section 4.2, a more powerful texture model is desirable to make the PEF more accurate.

In addition, the controlled experimental situation eliminates other considerations such as inconstant backgrounds, moving cameras and severe change of weather. To provide a general and really automatic approach to capturing human motion in unconstrained environments, much work remains to be done.

7. Conclusion

We have presented our work on the kinematics-based tracking of human walking parallel to the image plane in monocular image sequences based on the human body model composed of truncated cones and a sphere. This paper has three main contributions. Firstly, the PEF combining both boundary and region information derives the accuracy from the former and robustness from the latter. Secondly, according to the tree-like structure of human body model, we have designed a hierarchical search strategy to decompose the parameter space so that location of the human global position and estimation of each joint angle can be done separately. And the decomposition also reduces the search space. Thirdly, a kinematics-based algorithm has been proposed to recursively refine the joint angles. The experimental results on real sequences of both indoor and outdoor scenes have shown the effectiveness of our method.

Acknowledgements

The authors would like to thank Dr M. Nixon from University of Southampton, UK, for their help with the SOTON gait database. This work is supported by NSFC (Grant No. 60105002, 60373046, 69825105, and 60121302), the Natural Science Foundation of Beijing (Grant No. 4031004), and the National 863 High-Tech R&D Program of China (Grant No. 2002AA117010-11 and 2002AA142100).

References

- [1] D. Gavrilu, The visual analysis of human movement: a survey, *Comput. Vis. Image Understand.* 73 (1) (1999) 82–98.
- [2] T.B. Moeslund, E. Granum, A survey of computer vision-based human motion capture, *Comput. Vis. Image Understand.* 81 (2001) 231–268.
- [3] H.J. Lee, Z. Chen, Determination of 3D human body posture from a single view, *Comput. Vis., Graphics, Image Process.* 30 (1985) 148–168.
- [4] D. Hogg, Model-based vision: a program to see a walking person, *Image Vis. Comput.* 1 (1) (1983) 5–20.
- [5] Q. Delamarre, O. Faugeras, 3D articulated models and multi-view tracking with physical forces, *Comput. Vis. Image Understand.* 81 (2001) 328–357.
- [6] C. Sminchisescu, B. Triggs, Covariance scaled sampling for monocular 3D body tracking, *Proceedings of International Conference on Computer Vision and Pattern Recognition, CVPR' 2001, Kauai, HI.*
- [7] R. Plankers, P. Fua, Articulated soft objects for video-based body modeling, *Proceedings of Ninth International Conference on Computer Vision (ICCV' 2001), Vancouver, Canada.*
- [8] J. Aggarwal, Q. Cai, Human motion analysis: a review, *Comput. Vis. Image Understand.* 73 (3) (1999) 428–440.
- [9] S. Wachter, H.H. Nagel, Tracking persons in monocular image sequences, *Comput. Vis. Image Understand.* 74 (3) (1999) 174–192.
- [10] Y. Yang, M. Levine, The background primal sketch: an approach for tracking moving objects, *Machine Vis. Appl.* 5 (1992) 17–34.
- [11] Q. Delamarre, O. Faugeras, 3D articulated models and multi-view tracking with silhouettes, *Proceedings of Seventh International Conference on Computer Vision (ICCV'99), Kerkyra, Greece.*
- [12] J.C. Cheng, J.M.F. Moura, Capture and representation of human walking in live video sequence, *IEEE Trans. Multimedia* 1 (2) (1999) 144–156.
- [13] T. Zhao, T.S. Wang, H.Y. Shum, Learning a highly structured motion for 3D human tracking, *Proceedings of Fifth Asian Conference on Computer Vision (ACCV'2002), Melbourne, Australia (2002).*
- [14] D.M. Gavrilu, L.S. Davis, A 3-D model-based tracking of humans in action: a multi-view approach, *Proceedings of International Conference on Computer Vision and Pattern Recognition, San Francisco, CA (1996) 73–80.*
- [15] F. Lerasle, G. Rives, M. Dhome, A. Yassine, Human body tracking by monocular vision, *Proceedings of Fourth European Conference on Computer Vision, Cambridge, England (1996) 518–527.*
- [16] J.D. Shutler, M.S. Nixon, C.J. Harris, Statistical gait recognition via temporal moments, *Proceedings of Fourth IEEE Southwest Symposium on Image Analysis and Interpretation (2000) 291–295.*
- [17] J. Foster, M. Nixon, A.P. Bennett, New area based metrics for gait recognition, *Proceedings of Third International Conference on Audio- and Video-Based Biometric Person Authentication (2001) 312–317.*
- [18] J.B. Hayfron-Acquah, M.S. Nixon, J.N. Carter, Automatic gait recognition by symmetry analysis, *Proceedings of Third International*

- Conference on Audio- and Video-Based Biometric Person Authentication (2001) 272–277.
- [19] L. Wang, W. Hu, T. Tan, Recent developments in human motion analysis, *Pattern Recogn* 36 (3) (2003) 585–601.
- [20] J. Segen, S. Pingali, A camera-based system for tracking people in real time, *Proceedings of International Conference on Pattern Recognition*, Vienna (1996) 63–67.
- [21] D.S. Jang, H.I. Choi, Active models for tracking moving objects, *Pattern Recogn* 33 (7) (2000) 1135–1146.
- [22] S. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, H. Wechsler, Tracking groups of people, *Comput. Vis. Image Understand.* 80 (1) (2000) 42–56.
- [23] W. Wren, A. Azarbayejani, T. Darrell, A. Pentland, Pfinder: real-time tracking of the human body, *IEEE Trans. Pattern Anal. Machine Intell.* 19 (7) (1997) 780–785.
- [24] M. Kass, A. Witkin, D. Terzopoulos, Snakes: active contour models, *Int. J. Comput. Vis.* (1988) 321–331.
- [25] K. Toyama, A. Blake, Probabilistic tracking in a metric space, *Proceedings of International Conference on Computer Vision* (2001).
- [26] G. Mori, J. Malik, Estimating human body configuration using shape context matching, *Proceedings of European Conference on Computer Vision* (2002).
- [27] H. Ning, L. Wang, W. Hu, T. Tan, Articulated model based people tracking using motion models, the Fourth IEEE International Conference on Multimodal Interfaces (2002).
- [28] G. Borgefors, Hierarchical chamfer matching: a parametric edge matching algorithm, *IEEE Trans. Pattern Anal. Machine Intell.* 10 (6) (1988) 849–865.
- [29] M. Isard, A. Blake, CONDENSATION—conditional density propagation for visual tracking, *Int. J. Comput. Vis.* 29 (1) (1998) 5–28.
- [30] H. Sidenbladh, M. Black, D. Fleet, Stochastic tracking of 3D human figures using 2D image motion, *Proceedings of European Conference on Computer Vision* (2000).
- [31] L. Wang, T. Tan, W. Hu, H. Ning, Automatic gait recognition based on statistical shape analysis, *IEEE Trans. Image Process.* 12 (9) (2003) 1120–1131.
- [32] Y. Huang, T.S. Huang, Model-based human body tracking, *Proceedings of International Conference on Pattern Recognition* (2002).
- [33] H. Ning, L. Wang, W. Hu, T. Tan, Model-based tracking of human walking in monocular video sequences, the Seventeenth IEEE Region 10 Technical Conference on Computers, Communication, Control and Power Engineering (2002).
- [34] L. Wang, H. Ning, T. Tan, W. Hu, Fusion of static and dynamic features of body biometrics for gait recognition, *Proceedings of International Conference on Computer Vision*, Nice, France vol. II (2003) 1449–1454.
- [35] D. Lowe, Fitting parameterized 3-D models to images, *IEEE Trans. Pattern Anal. Machine Intell.* 13 (1991) 441–450.
- [36] D. Comaniciu, V. Ramesh, P. Meer, Real-time tracking of non-rigid objects using mean shift, *Proceedings of International Conference on Computer Vision and Pattern Recognition* (2000).
- [37] L. Lee, W.E.L. Grimson, Gait appearance for recognition, *ECCV Workshop on Biometric Authentication* (2002).
- [38] C.Y. Yam, M.S. Nixon, J.N. Carter, Gait recognition by walking and running: a model-based approach, *Proceedings of Asia Conference on Computer Vision*, Melbourne, Australia (2002) 1–6.
- [39] R. Tanawongsuwan, A. Bobick, Gait recognition from time-normalized joint-angle trajectories in the walking plane, *Proceedings of International Conference on Computer Vision and Pattern Recognition (II)* (2001) 726–731.