



PERGAMON

Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE @ DIRECT®

Pattern Recognition 37 (2004) 1423–1440

**PATTERN
RECOGNITION**

THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

People tracking based on motion model and motion constraints with automatic initialization

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

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, People's Republic of China

Received 19 December 2002; received in revised form 15 December 2003; accepted 5 January 2004

Abstract

Human motion analysis is currently one of the most active research topics in computer vision. This paper presents a model-based approach to recovering motion parameters of walking people from monocular image sequences in a CONDENSATION framework. From the semi-automatically acquired training data, we learn a motion model represented as Gaussian distributions, and explore motion constraints by considering the dependency of motion parameters and represent them as conditional distributions. Then both of them are integrated into a dynamic model to concentrate factored sampling in the areas of the state-space with most posterior information. To measure the observation density with accuracy and robustness, a pose evaluation function (PEF) combining both boundary and region information is proposed. The function is modeled with a radial term to improve the efficiency of the factored sampling. We also address the issue of automatic acquisition of initial model pose and recovery from severe failures. A large number of experiments carried out in both indoor and outdoor scenes demonstrate that the proposed approach works well

© 2004 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.

Keywords: Model-based human tracking; Motion model; Motion constraints; Initialization; CONDENSATION; Gaussian distribution

1. Introduction

The growing interest in human motion analysis is strongly motivated by recent improvements in computer vision, the availability of low-cost hardware such as video cameras and a variety of new promising applications such as personal identification and visual surveillance. It aims to automatically estimate the motion of a person or a body part from monocular or multi-view video images. Generally, the main motion parameters that should be estimated in each frame include orientation and translation of the human body and angles of all joints. Then the temporal information of motion parameters can be used in many applications 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, and video annotation [1,2,24].

It is true that a generic approach capable of capturing unconstrained human motions in real scenes would increase the flexibility of these applications. However, due to the complex nature of human body and real scenes, motion capture from image sequences is a very difficult task. It involves a large number of hard issues such as detection of the moving human in cluttered background with changing brightness, occlusion and self-occlusion, singular poses in kinematics, and a large number of degree of freedom (DOFs) (more than 30 for full body) [5]. Here we focus on a constrained problem: locating and tracking people walking along a line from monocular image sequences. Although the problem is constrained (so that it is more tractable), it is of great relevance to many practical applications such as visual

* Corresponding author. Tel.: +86-10-62647441; fax: +86-10-62551993.

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).

surveillance of a specific site where cameras may be installed in a desirable configuration.

Our ultimate goal is to acquire motion parameters including position, orientation and joint angles for gait recognition. Previous work on gait recognition mainly adopted low-level information such as silhouette [3,4]. These methods often overlooked high-level information, e.g., temporal features of joint angles, which are much more essential to reflect the dynamics of gait motion. To achieve more accurate results, we intend to use high-level information or combine both low-level and high-level features for gait recognition. As an important step, this paper focuses on acquisition of motion parameters through a model-based tracking approach.

2. Related work

Human motion analysis is an active and growing research area [1,2,27,28]. Here we briefly outline model-based tracking methods and review some previous work on human body models, motion models and search strategies in order to put our work in context.

Model-based tracking approach is very popular in human motion analysis recently. In such an approach, a geometric human body model is represented by a number of joints and sticks that connect each other according to the human body structure. The sticks, “fresh” surrounding them, and skin texture may be represented in different ways depending on what the systems need. The concrete equivalent representation of a human body model is a state vector that indicates the current pose of the tracked human. The pose of the subject is a point in the state space while corresponding to many points in the 2D image space. So the essence of model-based tracking is to relate the image data to pose data. The general approach is known as *analysis-by-synthesis*, and is used in a *predict-match-update* style [27]. With such an approach, the pose of the model for the next frame is first predicted according to prior knowledge and motion history. Then, the predicted model is synthesized and projected into the image plane for comparison with the image data. A specific PEF is needed to measure the similarity between them. According to different search strategies, this is done either recursively or using sampling techniques until the correct pose is finally found and used to update the model. With the exception, pose estimation in the first frame needs to be handled specially.

Model-based tracking has three major advantages. It can obtain detailed and accurate motion data that can be used in many real applications. It has the ability to cope with occlusion and self-occlusion. Finally, it enables the prior knowledge (such as human body structure, motion constraints and motion model) to be incorporated very easily. Generally, model-based tracking involves three main tasks: construction of human body model, representation of prior knowledge of motion model and motion constraints, and

search strategy (e.g., prediction). In the following, we briefly review the previous work in these areas.

As mentioned above, motion capture involves many hard issues. To alleviate these difficulties, a variety of human body models and motion models were introduced as priors in previous work. As far as human body models are concerned, they vary widely in the levels of details. In earlier research, simple stick figure models were frequently used [6,29], in which body parts are represented by sticks connected by joints. More complex volumetric human models, such as cylinder [7,8,30], truncated cone [9,10] and super-quadrics [11], were used in later work. Recently, Plankers and Fua [12] presented a hierarchical human model to achieve more accurate results, which included four levels: skeleton, ellipsoid meatballs simulating tissues and fats, polygonal surface representing skin, and shaded rendering. As a general rule, the more complex the human body model, the more accurate tracking results may be expected but at the expense of higher computational complexity.

Also, motion models of body limbs and joints are widely used in the tracking process. They serve as prior knowledge to predict motion parameters [5,14], to interpret and recognize human dynamics [15], or to constrain the estimation of low-level image measurements [13]. For instance, Bregler [15] decomposed human dynamics into multiple abstractions, and represented the high-level abstraction by hidden Markov model (HMM) as successive phases of simple movements. This representation was used for both tracking and recognition. Zhao [5] trained a highly structured motion model for ballet dancing under the minimum description length (MDL) paradigm. This motion model is similar to a finite state machine (FSM). The popular method, multivariate principal component analysis (MPCA), was recently used to train a walking model in Sidenbladh et al. [13]. Similarly, Ong and Gong [19] employed the hierarchical PCA to learn their motion model that was represented by the matrixes of transition probabilities between different subspaces in a global eigenspace and by the matrix of that between global eigenspaces. Unlike these methods, we learn a motion model from semi-automatically acquired training examples and represent it as Gaussian distributions.

Pose estimation in a high-dimensional body configuration space is intrinsically difficult, so, in previous work, search strategies were often carefully designed to reduce the solution space. Generally, four main classes of search strategies exist: kinematics, Taylor models, Kalman filtering 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 [9]. Taylor models incrementally improve an existing estimation, using differentials of motion parameters with respect to the observation to predict better search directions [21]. It at least finds local minima, but cannot guarantee global optimality. As a recursive linear estimator, Kalman filtering

can thoroughly deal with the tracking of shape and position over time in the relatively clutter-free case in which the density of the motion parameters can satisfactorily be modeled as Gaussian [8,20]. To handle clutter that causes the density of motion parameters to be multi-modal and non-Gaussian, methods of stochastic sampling, such as Markov Chain Monte Carlo [22], Genetic algorithms and CONDENSATION [5,17,18,23], are designed to represent simultaneous alternative hypotheses. Among the stochastic sampling methods in visual tracking, CONDENSATION is perhaps the most popular. And it is therefore also used as the tracking framework in this paper.

3. Outline of our approach

In this paper, we present an effective approach to tracking walking human based on both body model and motion model in a CONDENSATION framework [17]. The CONDENSATION framework is very attractive because it can handle clutter and fusion of information. Fig. 1 gives the framework of our approach. In tracking, we maintain, at successive time-steps, a sample set of poses that are 12-dimensional vectors described in Section 4. The sample set is derived either from the tracking result of previous frame, or from a specific initialization process when it comes to the first frame. For each new frame, the sample set is subjected to the predictive steps. First, samples undergo drift according to previous pose, motion model and motion constraints. The second predictive step, diffusion, is random and the drifted samples may split. Our dynamic model directs the predictive steps. After prediction our PEF measures the similarities between the image data and the projected human body model with diffused poses. And the posterior mean pose of the tracked people can be generated from the sample set by weighting with the similarities. In Fig. 1, the left dashed rectangle indicates the dynamic model and the right one is the measurement. The tracking results will finally be used for gait recognition that is part of our future work.

In this paper, the human body model is represented by articulated truncated cones and a sphere as a trade-off between accuracy and complexity. Our motion model that is different from previous ones such as Sidenbladh et al. [13] is learnt from semi-automatically acquired training data. Using the motion model, we explore the dependency between the shoulder-elbow joint and the thigh-knee joint to discover and describe the motion constraints. This is different from previous work that usually only considered constraints of intervals of joint angles. These constraints, together with the motion model, are integrated into the dynamic model to concentrate the factored sampling areas. Our PEF combines both boundary and region information that makes it accurate and robust. We model it with a radial term to improve the efficiency of factored sampling. Also, the tracker can automatically initialize the sample set when tracking the first frame and recovering from severe failures.

This paper is an extended version of our previous work described in [32]. The major modification lies in the motion model, initialization, dynamical model, experiments on real-world outdoor data, analysis of some failure modes, and gait recognition. The main contributions of this paper are as follows:

- Compact and effective motion model and motion constraints are automatically learnt that enforce the dynamic model for the CONDENSATION algorithm and result in a low computational cost.
- An accurate and robust pose evaluation function is proposed that highly reduces the size of sample set required by the CONDENSATION framework.
- Compared with previous work of initialization [5,13,8] that is usually done manually, ours is automatic. Cheng [14] also 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 a high computational cost. In contrast, our

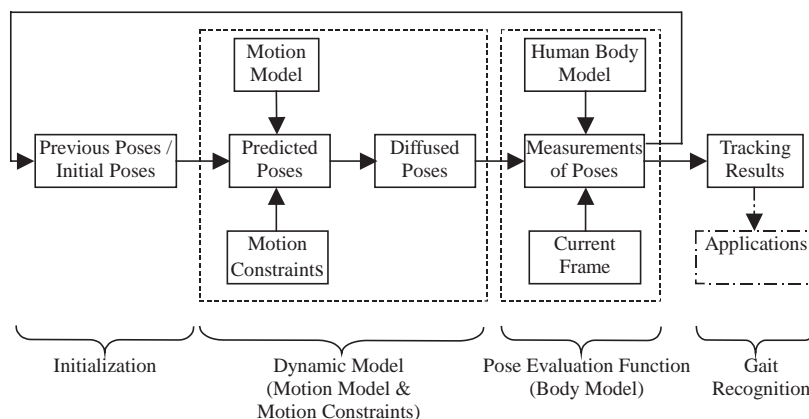


Fig. 1. Framework of our approach.

method, employing the spatio-temporal information to avoid computing the PEF, is much faster.

- Dynamic features of individual gait (i.e., joint-angle trajectories) are used to identify people and the recognition rate seems higher than our other work using static features [33].

The remainder of this paper is arranged as follows. Section 4 describes the human body model and its state vector; Section 5 details the learning process of the motion model and the motion constraints; Section 6 presents our approach to tracking in the CONDENSATION framework; Section 7 gives some experimental results and discussions; and Section 8 contains concluding remarks and outlines future work.

4. Human body model

Our human body model, similar to [13,8], 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 that is represented by a sphere. They are connected to others at joints, the angles of which 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.

The above human body model in its general form has 34 DOFs: 3 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 along a line when the camera is installed in a desirable configuration (For convenience, we assume that people walk parallel to the image plane. With little modification, our approach can also be applied to other fixed directions.), 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 when the upper torso moves along a line. Furthermore, each joint has thus only one DOF. Accordingly, 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 position and posture 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.

In model-based tracking, we need to synthesize and project the body model into 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 camera coordinate of each point in the body model and transform it to the image coordinate. To locate the positions of model parts in the camera coordinate, 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 (the joint position). We represent the human body model as a kinematical tree, with the torso at its root, to order the transformations between the local coordinate frames of different parts. Therefore, the camera coordinate of each part is 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. The geometrical optics is modeled as a pinhole camera with a transformation matrix T such that $X_i = T \bullet X_c$, where X_i and X_c are image and camera coordinates of a point on the human body respectively (See [13] for more information).

5. Learning motion model and motion constraints

A motion model, encoding the dynamics of the human body, can be used in tracking to greatly reduce the computational cost while achieving better results. As a highly constrained activity, the gait patterns of human walking are symmetric, periodical and of little variation in a wide range of people [16]. So it is relatively easy to learn a compact

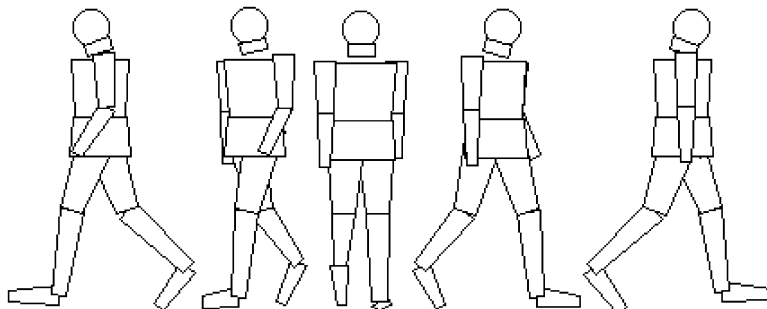


Fig. 2. Human body model projected into the image plane from 5 viewing angles.

and effective motion model for human gait from limited training data. In this paper, our motion model for human gait (hereafter referred to as motion model) is learnt from semi-automatically acquired training examples and formulated as Gaussian distributions. Also, the dependency of joint angles is analyzed to explore the motion constraints that, together with the motion model, are integrated into the dynamical model to focus on the heavy weighted samples in the CONDENSATION framework.

5.1. Learning motion model

In the learning process, training data (9 examples from 5 different subjects) corresponding to the motion parameters $P = \{x, y, \theta_1, \theta_2, \dots, \theta_{10}\}$ were semi-automatically acquired by a specially designed software. Several feature points in each frame are marked manually and the motion parameters derived from these features are computed and analyzed automatically. Some acquired data are illustrated in Fig. 3(a) which reveals that the temporal curves have different periods and phases. Therefore the walking cycles in each training example must be rescaled to the same length and aligned to the same phase before learning the motion model.

To compute the period and phase of a sequence, we define the correlation function *corr* with respect to two matrices $A_{m \times n}, B_{m \times n}$ having the same size.

$$\text{corr}_i(A, B) = \frac{\|A \bullet T_B(i)\|}{\|A\| \|T_B(i)\|}, \quad i = 1, 2, \dots, m, \quad (1)$$

where $A \bullet B$ returns the matrix whose elements are products of the corresponding elements in A and B . $T_B(i)$ removes the first $i - 1$ rows of B and add $i - 1$ rows of zeros to the end of B . When A and B have different rows, we add enough zeros to end of the matrix that has less rows to make their rows equal.

We form matrix A'_i for each training example i with row index indicating time step and column index indicating motion parameters, where $i = 1 \dots m$. Then the period T_i of the i th example is computed from the cross-correlation $\text{corr}(A'_i, A'_i)$. The interval between two dominant peaks is chosen as the period. To rescale the walking cycle to the same length T (in this paper $T = 100$), the B-spline interpolation algorithm is applied to the example A'_i with the scalar $a_i = T/T_i$. Given that A'_i is rescaled to A_i , a specific one in A_i ($i = 1 \dots m$), e.g. A_1 , is selected as the reference. Then the phase b_i of each example A_i relative to the reference example A_1 is indicated by the predominant peak in the cross-correlation $\text{corr}(A_i, A_1)$. In all,

$$B_i(t) = A'_i(t/a_i + b_i), \quad i = 1 \dots m \quad (2)$$

are the normalized examples with the same period and phase. The segments $B_i(1 : T), B_i(T + 1 : 2T), \dots, i = 1 \dots m$, renamed as W_j with $j = 1 \dots n$, are exactly all of the

normalized walking cycles. Then our motion model is empirically represented as Gaussian distributions $G_{k,t}(u_{k,t}, \sigma_{k,t}^2)$ for each joint angle k ($k = 1 \dots 10$) at any phase t ($t = 1 \dots T$) in the walking cycle with

$$u_{k,t} = \frac{1}{n} \sum_{j=1}^n W_j(t, k), \quad (3)$$

$$\sigma_{k,t} = \sqrt{\frac{1}{n} \sum_{j=1}^n (W_j(t, k) - u_{k,t})^2}. \quad (4)$$

Fig. 3(b) and (c) are temporal models of joint angles of left thigh and left knee. Although learnt from limited data, they correspond very well to Murray's results in medical analysis [16]. The learning and representation of our motion model are compact, and it shows great effectiveness in estimation of the prior distribution of initial pose and in prediction of new pose for the next frame. Then a question arises: is it reasonable to assume the Gaussian distributions for the gait model? According to experience, the variation of joint k at the phase t in a walking cycle should be Gaussian. To verify this assumption, we randomly select the joint k and phase t in a walking cycle, and histogram the data set $\{W_j(t, k), j = 1 \dots n\}$. Fig. 4 gives two examples. To make it clear, the Gaussian distributions are also plotted with the mean and deviation computed by Eqs. (3) and (4). It can be seen that the Gaussian curves fit the histogram very well. This experiment basically verifies that our assumption is reasonable. However, we still need to capture more training data to acquire a more accurate probabilistic representation of the motion model.

5.2. Motion constraints

Although the motion model describes the basic pattern of walking, it does not contain all information about walking. Therefore, we derive motion constraints from training data by further exploring the dependency of neighboring joints: shoulder and elbow, thigh and knee, knee and ankle. Obviously, in a walking activity, the movements of the lower arm and the upper arm are correlated and regular, so the shoulder joint and the elbow joint are not independent. We assume that the lower arm is driven by the upper arm, and accordingly the elbow joint is determined by the shoulder joint except for some noise. So the motion constraint of the elbow joint can be approximated by the conditional distribution $p(\theta_e | \theta_s)$, where θ_e and θ_s are the joint angles of the elbow and the shoulder respectively. Using the training data in the previous subsection, the distribution can be easily computed by the following procedure. From each walking cycle W_i ($i = 1 \dots n$), a series of pairs of the shoulder and elbow joint angles $(\theta_{is}(t), \theta_{ie}(t))$ are formed as the time t varies from 1 to T . We classify all pairs according to their first element, i.e., pairs having identical first element are

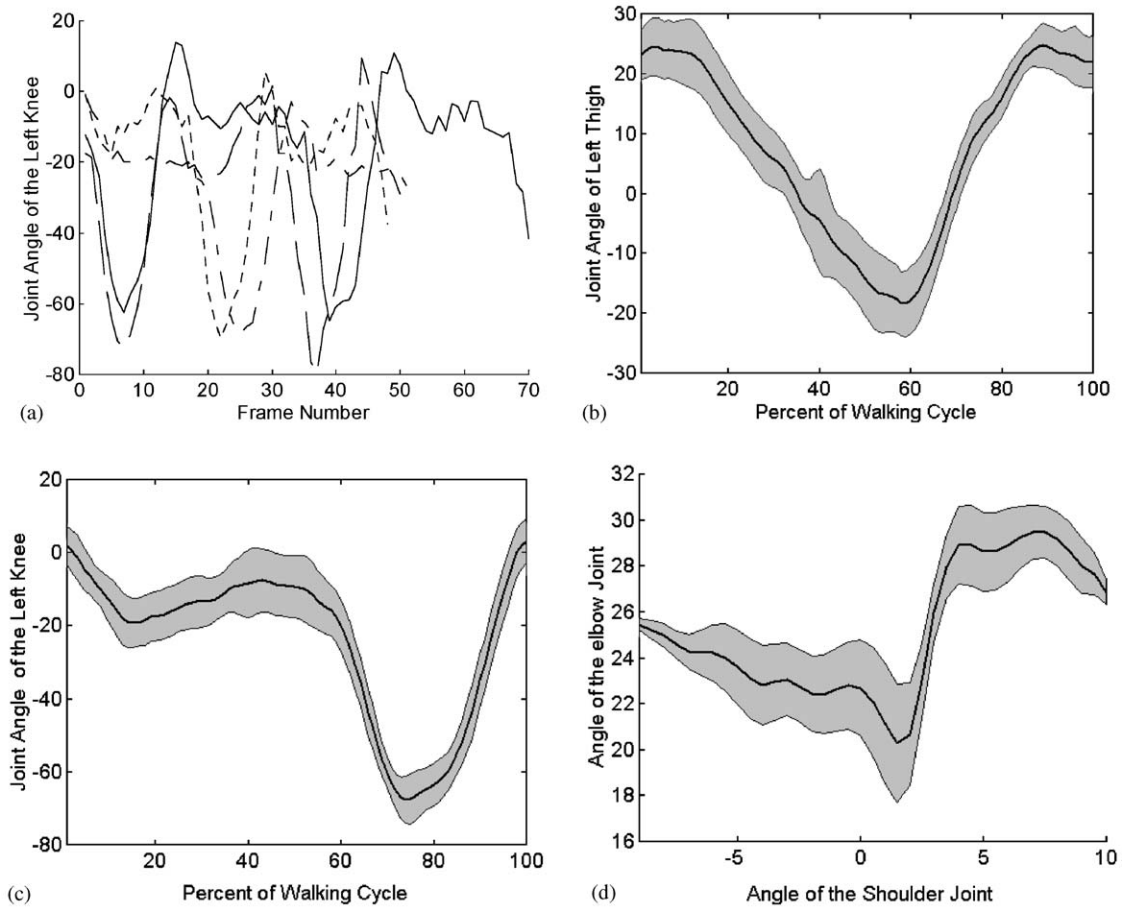


Fig. 3. Motion model and motion constraints, (a) joint angles of the left knee of four different people walking with various periods and phases; (b) and (c) temporal models of joint angles of left thigh and left knee during a walking cycle; (d) motion constraints of the elbow joint. In (b), (c) and (d), the dark lines and the shaded areas indicate the mean and standard deviation of the corresponding distribution, respectively.

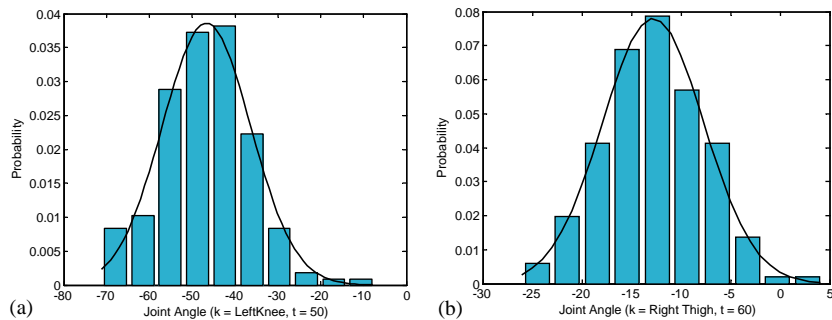


Fig. 4. Examples of the distributions of a joint at a fixed phase in the walking cycle, (a) the distribution of the joint of the left knee at phase 50, (b) the distribution of the joint of the right thigh at phase 60. In (a) and (b), the histograms are derived from the training data and the thick curves are Gaussian functions.

assigned to the same class. Provided that (θ_s, \bullet) is a class for shoulder joint angle θ_s including K pairs (θ_s, θ_e^k) , $k = 1 \dots K$, the conditional distribution $p(\theta_e | \theta_s)$ is represented by a Gaussian distribution $G(u, \sigma^2)$ where u and σ are the mean and standard deviation of θ_e^k , $k = 1 \dots K$. Fig. 3(d) gives the motion constraint for the elbow joint. The motion constraints for the knee and ankle joint are learnt in the same way. Here, Gaussian representation is assumed for simplicity which seems to work well. However, it needs further analysis in future work.

We also derive intervals of valid value for each motion parameter from training data by specifying its maximal and minimal value. All the generated samples are constrained in their associated intervals by setting the over-set values to its minimum or maximum.

6. Tracking

The main task here is to relate the image data to the pose vector $P = \{x, y, \theta_1, \theta_2, \dots, \theta_{10}\}$ defined in Section 4. Since the articulated human body model is naturally formulated as a tree-like structure, a hierarchical estimation, i.e., locating the global position and tracking each limb separately, is suitable here, especially when the total number of parameters is large. Additionally, this approach to decomposing the parameter space is strongly supported by other reasons as follows. Firstly, the global position (x, y) is much more significant than other joint angles with respect to the PEF (see Section 6.3), so it can be estimated separately with other motion parameters fixed. Secondly, joint parameters are greatly dependent on the global position (x, y) . In detail, a slight deviation of the global position often cause the joint parameters to drastically deviate from their real values when maximizing the PEF, and the result is that a large weight is assigned to the actually unimportant sample. So the global position should be located ahead of sampling the joint angles. Thirdly, locating the optimal pose in a high-dimensional state space, e.g., 12 DOFs in this paper, is intrinsically difficult. Decomposition will effectively simplify this problem. Finally, because sometimes 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 accordingly needs a larger sample set.

Given the above considerations, we first predict the global position from the centroid of the detected moving human and then refine it by searching the neighborhood of the predicted position. Each limb is tracked under the CONDENSATION framework [17]. As a popular method in visual tracking, the CONDENSATION algorithm uses learnt dynamical models, together with visual observations, to propagate the random sample set over time. Instead of computing a single most likely value, it evaluates the posterior distribution by factored sampling and thus can represent simultaneous alternative hypotheses. Therefore, it is more robust than Kalman filter, a Gaussian-based and unimodal method.

Another advantage of the CONDENSATION framework is that it can easily handle fusion of information, especially temporal fusion, in a principled manner. Later, we can see that the information of observation, prior knowledge of motion model and motion constraints are all straightforwardly fused by the density propagation rule to derive the posterior distribution.

The rule of state density propagation over time is [17]

$$p(x_t | Z_t) = k_t p(z_t | x_t) \int_{x_{t-1}} p(x_t | x_{t-1}) \times p(x_{t-1} | Z_{t-1}) dx_{t-1}, \quad (5)$$

where x_t are the motion parameters at time t , $Z_t = (z_1, z_2, \dots, z_t)$ is the image sequence up to time t , and k_t is a normalization constant independent of x_t . According to this rule, the posterior distribution of $p(x_t | Z_t)$ can be derived from the posterior $p(x_{t-1} | Z_{t-1})$ at the previous time step and three other components: the prior distribution $p(x_0)$ at time 0, i.e., the initialization, the dynamical model $p(x_t | x_{t-1})$ to predict the motion parameters x_t by drifting and diffusing x_{t-1} , and the observation density $p(z_t | x_t)$ computed from the PEF. They are, respectively, detailed in the following subsections.

6.1. Initialization

Initialization is concerned with the initial pose of a subject in capturing human motion. Most 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 uniform distribution [5,8,13]. There is also some work presenting basic methods to overcome this hard but important issue. For instance, Sminchisescu and Triggs [25] adopted a hierarchical three-step process to obtain the initial pose and human body model, but it involved a difficult problem of finding the correspondences between the given 3D model and the 2D image. Hoshino and Saito [20] estimated the initial body pose by extracting the centerline of each body parts from 2D input image, but they did not provide any methods to accurately extract centerlines. Obviously this is very challenging in real scenes. In addition, Cheng [14], as mentioned in Section 3, also provided an initialization method that evaluates the cost function many times, leading to a high computational cost. Unlike previous work on initialization that attempts to roughly estimate the pose from a single frame, we accomplish it using spatio-temporal information of the first N frames. Thus our approach is more robust, and most importantly, it also achieves real-time speed by avoiding evaluating the cost function. In what follows, we describe the initialization procedure that includes a learning process and an estimation process.

In the learning process, the moving human in each frame in the training data is detected by subtracting the background image and extracted edges using the Sobel operator. And

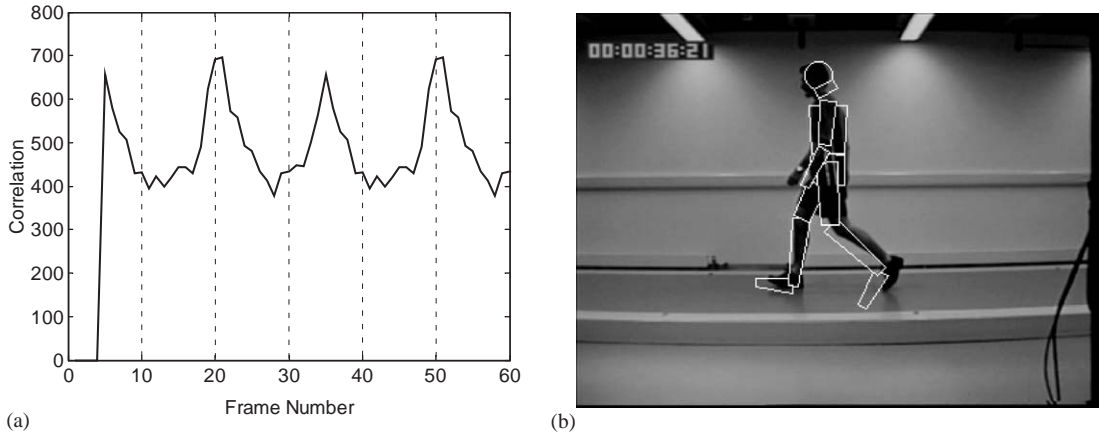


Fig. 5. Example of initialization (for frame 19 in sequence mp2). (a) cross-correlation between the short sequence (frame 15–19 in sequence mp2) and two concatenated walking cycles; (b) projection of the human body model with the initial pose.

then, the moving area is clipped and normalized to the same size. Similar to the preprocessing of learning motion model, the normalized examples are adjusted to the same phase and their periods are rescaled to the same length (here the length is $T_{awc} = 30$). Also, n normalized walking cycles V_j with $j=1 \dots n$, are segmented from the m preprocessed examples. We use the average walking cycle

$$V = \frac{1}{n} \sum_{j=1}^n V_j \quad (6)$$

as the reference cycle.

The estimation process begins with the same preprocessing of human detection, edge extraction and normalization as that in the learning process. The first N frames $v(N < T_{awc})$ are then located in the reference cycle by searching the major peak in the cross-correlation $corr(V, v)$. Referring the location (assumed to be t) to the motion model, the pose of the last frame in v is roughly estimated as a 10-dimensional vector $(u_{3,t}, u_{4,t}, \dots, u_{12,t})$. Accordingly, the prior distribution for tracking is the Gaussian distribution $G((u_{3,t}, u_{4,t}, \dots, u_{12,t}), \sigma_t^2 I_{10})$ where I_{10} is a 10×10 identity matrix and $\sigma_t^2 = (\sigma_{3,t}^2, \sigma_{4,t}^2, \dots, \sigma_{10,t}^2)$.

Fig. 5 illustrates the estimation process. To estimate the initial pose for frame 19 in sequence mp2, the short sequence (frame 15–19) is used to compute the cross-correlation at all displacements between the short sequence and the average walking cycle. Here two average walking cycles are concatenated to make the result more accurate. Assuming the predominant peak over all shifts locates at t ($t = 21$, see Fig. 5(a)), we may derive the phase in the motion model for frame 19 by $p = t \times T/T_{awc}$ where T is defined in Section 5.1. Then the motion model at phase p indicates the initial pose. In Fig. 5(b), the human body model with initial pose is projected to the real image data. The result shows that,

although there are some errors at the left arm and the left leg, the initialization as a whole is very close to the true pose, demonstrating the effectiveness of our automatic initialization procedure.

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 PEF reaches a predefined threshold), the tracker will stop for N frames and reinitialize using the spatio-temporal information derived from such N frames to estimate the current pose. However, it is worthwhile to mention that the real-time speed and robustness of the initialization and bootstrap is at the expense of the first $N - 1$ frames in which tracking is stopped.

6.2. Dynamic model

The dynamic model is often carefully designed to improve the efficiency of factored sampling. The idea is to concentrate the samples in the areas of the state space containing most information about the posterior. The desired effect is to avoid as far as possible generating samples that have low weights, since they contribute little to the posterior. In this paper, the learnt motion model served as prior is integrated into the dynamic model to achieve efficiency of sampling. In detail, with the assumption that the Gaussian distributions at different phases in the motion model are independent, at any time instant t the i th motion parameter $\theta_{i,t}$ satisfies the dynamic model

$$p(\theta_{i,t} | \theta_{i,t-1}) = G(\alpha u_{i,t} + \beta u_{i,t-1} + \gamma \theta_{i,t-1}, \lambda((\alpha \sigma_{i,t})^2 + (\beta \sigma_{i,t-1})^2)), \quad (7)$$

where G is a Gaussian distribution, and $\alpha + \beta + \gamma = 1$ makes the drifting of $\theta_{i,t}$ not only from the tracking history $\theta_{i,t-1}$ but also from the motion model, and λ is a scalar that is

often set to 1. But when the gait of the tracked person is very normal, a smaller λ is expected to restrict the factored sampling more effectively to portions of the parameter space that are most likely to correspond to human motion. $u_{i,t}$ and $\sigma_{i,t}$ are defined in Section 5.1.

This dynamic model is generally sufficient for all motion parameters, but motion constraints can further concentrate the samples for motion parameters: elbow, knee and ankle joint. For instance, after the shoulder joint $\theta_{s,t}$ is sampled, sample positions generated from the conditional distribution $p(\theta_{e,t} | \theta_{s,t})$ (see Section 5.2) for the elbow joint $\theta_{e,t}$ also contain much information. So a mixed-state CONDENSATION [18] can be included in the factored sampling scheme by choosing with probability q to generate samples from the dynamic model (7) and with probability $1 - q$ to generate samples from the conditional distribution $p(\theta_{e,t} | \theta_{s,t})$, i.e., $\theta_{e,t}$ satisfies the dynamic model

$$p(\theta_{e,t} | \theta_{e,t-1}, \theta_{s,t}) = qG(\alpha u_{e,t} + \beta u_{e,t-1} + \gamma \theta_{e,t-1}, \lambda((\alpha \sigma_{i,t})^2 + (\beta \sigma_{i,t-1})^2)) + (1 - q)p(\theta_{e,t} | \theta_{s,t}), \quad (8)$$

where $\alpha, \beta, \gamma, \lambda$ are defined as above. Equations similar to (8) can also be provided for knee and ankle joints.

The core of the CONDENSATION algorithm is factored sampling using the dynamic models. But the general algorithm is subject to the main drawback that sample locations are determined purely by prediction from past observation. Re-sampling CONDENSATION [26], relying on the variation within the spatial model being separated into pseudo-independent components, overcame this drawback to some extent. However, the assumption of pseudo-independency can hardly be satisfied in our situation. Therefore, we introduce a feedback in the factored sampling to reduce the dependency on the prediction from the past observation. In detail, the weights of the samples in the old sample set are updated dynamically according to their importance measured by the PEF in the current frame. If a sample selected from the old sample set, after drifting and diffusing, has a weight change Δw in the current frame with respect to its old weight w , w is then altered (to w') to adjust the possibility of the next selection according to the formula $w' = w + \eta \Delta w$, where $0 \leq \eta \leq 1$ is a small scalar that determines the influence of the current observation on the factored sampling. Obviously, the feedback algorithm makes the sample location dependent on the prediction not only from the past observation but also from the current frame. And the generated samples will gradually approach the current observation. It should also be noticed that the extended feedback algorithm leads to no computational overhead over the standard CONDENSATION algorithm except for the simple calculation of the adjustment formula. Fig. 6 gives an example of factored sampling for the left lower limb.

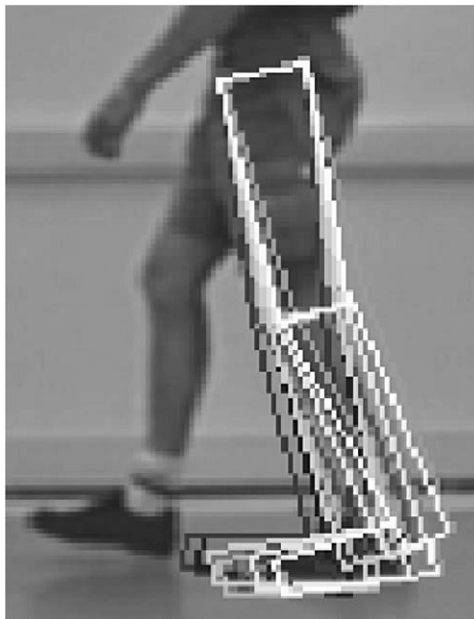


Fig. 6. Factored sampling for the left lower limb (only 10 samples are randomly drawn from the sample set).

6.3. Pose evaluation function

The PEF reveals the observation density $p(z_t | x_t)$ of an image z_t given that the human model has the posture x_t at time t . To match the image z_t with the generative model, the model must be projected into the image plane. Furthermore, detection based on background subtraction and Sobel operator are applied successively to the image z_t to acquire both region and boundary information. In general, boundary information improves the localization, whereas region information stabilizes the tracking because more of the image information is used. Therefore, we combine them in the PEF by computing boundary matching error and region matching error so as to achieve both accuracy and robustness.

Fig. 7 shows the procedure of computing boundary matching error that is similar to the chamfer distance. For each pixel p_i in the boundary of the projected human model, 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 that direction is desired. Given that q_i is the corresponding pixel and that F_i stands for the vector $\vec{p_i q_i}$, the matching error of pixel p_i to q_i can be measured as, the norm $\|F_i\|$. Then the average of the matching errors of all pixels in the boundary of the projected human model is defined as the boundary matching error

$$E_b = \frac{1}{N} \sum_{i=1}^N \|F_i\|, \quad (9)$$

where N is the number of the pixels in the boundary.

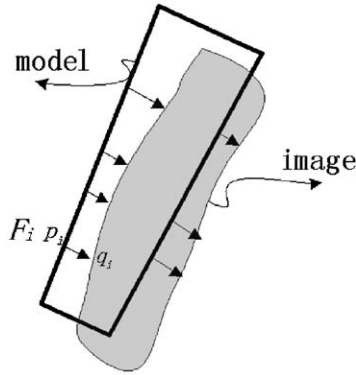


Fig. 7. Measuring the boundary matching error.

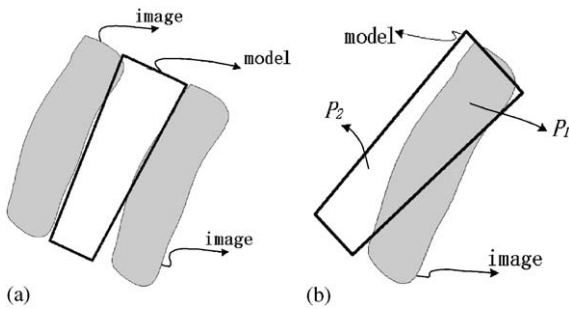


Fig. 8. Illustration of the necessity of simultaneous boundary and region matching, (a) a typical ambiguity: a model part falls into the gap between two body parts in the image, (b) measuring region matching error.

In general, the boundary matching error can properly measure the similarity between the human model and image data, but it is insufficient under certain circumstances. A typical example is given in Fig. 8(a), where a model part falls into the gap between two body parts in the edge image. Although it is obviously badly fitted, the model part may have a small boundary matching error. To avoid such ambiguities, region information is further considered in our approach. Fig. 8(b) illustrates the region matching. Here the region of the projected human model that is fitted into the image data is divided into two parts: P_1 is the region overlapped with the image data and P_2 stands for the rest. Then the matching error with respect to the region information is defined by

$$E_r = |P_2| / (|P_1| + |P_2|), \tag{10}$$

where $|P_i|$, ($i = 1, 2$) is the area, i.e., the number of pixels in the corresponding region.

Both boundary and region matching errors are combined into the PEF that is modeled in terms of a robust radial term $\rho_i(s, \sigma) = v e^{-s/\sigma^2}$ [11]

$$S(P) = v e^{-(\alpha \times E_b + (1-\alpha) \times E_r) / \sigma^2}, \tag{11}$$

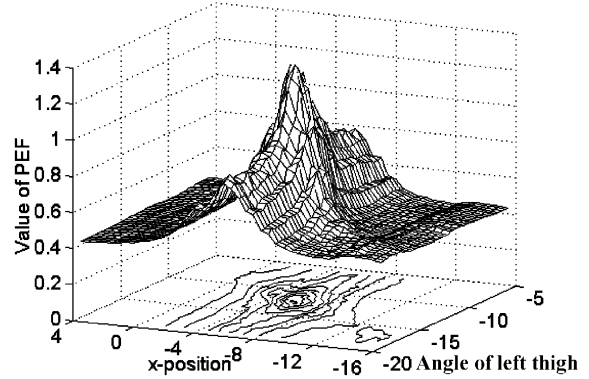


Fig. 9. The curve of the PEF with the global position x and the joint angle of the left thigh changing smoothly and other parameters remaining constant. Also shown is the contour of the function. The curve is shifted up by 0.4 to make the contour clearer.

where $P = \{x, y, \theta_1, \theta_2, \dots, \theta_{10}\}$ is the pose vector, and α is a scalar to adjust the weights of E_b and E_r . Apart from its robustness, the radial term can improve the efficiency of factored sampling because it assigns heavier weights to important samples and reduces the weights of insignificant ones. A smaller σ will magnify this effect, but it also makes the curve of the PEF peakier, which leads to a lower survival rate of samples. And then the needed number of samples increases. Therefore, σ must be carefully selected. As far as α is concerned, a bigger value is preferred for the upper limbs to diminish 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 also frequently occlude each other, and therefore the region information is of relatively little importance.

Fig. 9 shows the effectiveness of the proposed PEF. Its curve is basically smooth and has no local maxima at the neighborhood of the global maximum. These two properties are very useful for optimization. Furthermore, according to the contour of the PEF and other experiments, we can conclude that the global position (x, y) is more significant than other joint angles with respect to the PEF. This is one of the reasons why the global position can be firstly determined with other parameters fixed.

Our PEF is insensitive to noise to some extent. It is true that the PEF is dependent on the boundary and motion detection that is sensitive to noise. But in tracking, each limb is considered as a whole. Although a part of the limb is affected by noise, the PEF can still realistically reveal the pose when the total limb is considered. In other words, the PEF can utilize the prior knowledge of body model to reduce the influence of noise to some extent. For example, in Fig. 10 the left upper arm and right leg are missed in the edge image (b) due to noise but the failure was recovered in tracking where each limb is tracked in its entirety. However, when the whole boundary of the limb cannot be detected, the PEF

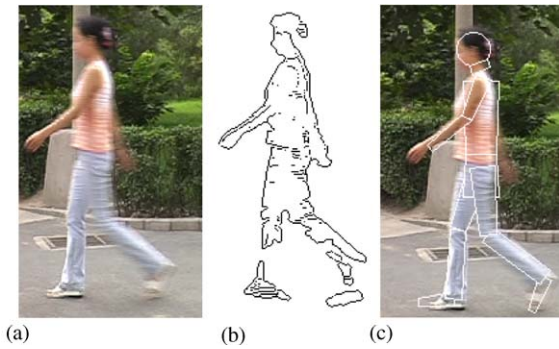


Fig. 10. Noise sensitivity for the PEF, (a) original image; (b) edge image where the left upper arm and right leg are missed due to noise, (c) the left upper arm and right leg are correctly tracked.

may fail that results in incorrect tracking (see frame 32 in Fig. 17 where the left arm is missed due to significant motion blur).

7. Experiments and discussions

To examine the tracking performance of the proposed algorithm, we conduct experiments on several persons having various shapes and walking characteristics in image sequences with low quality and significant self-occlusion. We restrict these persons to walk in the image plane. The approach is implemented with the MATLAB on a personal desktop workstation. The tracked sequences are selected both from the training data and from new instances. For each new instance, the tracker needs 300 state samples for

the upper limbs and 100 state samples for the lower limbs. Comparatively, due to the accuracy of the motion model, each sequence from the training data only requires 100 and 50 state samples for the upper and lower limbs, respectively. We use the first 15 frames of each sequence to automatically initialize the tracker.

7.1. 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 [31]. The database includes six subjects and four sequences for 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. It is noted that the training examples are all selected from this database. The outdoor sequences are captured at the same frame rate by a digital camera (Panasonic Nv-Dx100EN) fixed on a tripod. 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 for each subject. Some samples are showed in Fig. 11. It can be seen that the background of the image sequences in the NLPR database is basically clean though they were captured in outdoor scenes. So we also captured some additional sequences in more complex real-world outdoor scenes to further test our approach.

7.2. Tracking results and discussions

Started automatically by the initialization procedure described in Section 6.1, the system tracks successfully in



Fig. 11. Samples in the NLPR gait database.

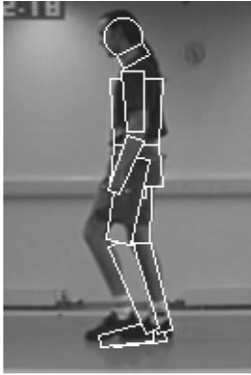


Fig. 12. Failure of tracking the left leg without considering region information.

the entire image sequences, though sometimes stopped and reinitialized for severe failures most likely due to occlusion, accumulated errors, or drastic image noise. Here the tracking results of three sequences from the training data (see Figs. 12 and 13) and two new instances (see Fig. 14) are showed. Due to space constraint, only the human areas clipped from the original image sequences are given. Some sequences include challenging configurations in which the two legs and thighs occlude each other severely (e.g. frame 27 in Fig. 13(a)), causing most part of one leg or thigh is unseen. These difficult 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 worthwhile to mention that sometimes the arms off the camera in these sequences were lost for the severe occlusion by the torso (see frame 19 in Fig. 13(c)). However, their motion parameters can usually be properly estimated using the motion model (see frame 27 in Fig. 14(b)) or using the symmetric value of the other arm.

In our earlier experiments, region information is not considered. So when the human legs are in the special posture showed in Fig. 12, the model part of left leg falls into the gap between the two legs and the tracking fails. In contrast, through the combination of the two matching errors, the tracking results are much more robust (see frame 29 in Fig. 13(b)). This example clearly illustrates the effectiveness of the proposed PEF.

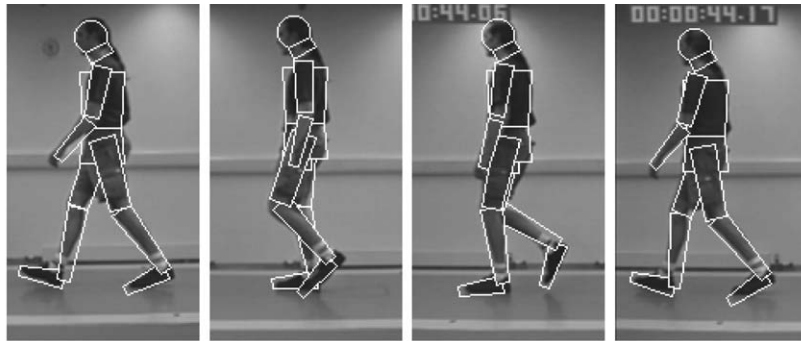
Also, further experiments are carried out to deeply analyze our approach. It is mentioned that tracking new sequences requires much more state samples than tracking sequences from the training data. The chief reason is that our motion model is only learnt from limited training data and cannot accurately represent the variations of all gait sequences, especially the abnormal ones. When encountered a novel instance, the deficiency of the motion model will reduce the accuracy of the prediction of the dynamic model. Fortu-

nately, increasing the state samples will compensate it. This is demonstrated by an experiment. In Fig. 15(a), when the sequence is included in the training data, the prediction is very close to the refined results. The good prediction, which also proves the effectiveness of the dynamic model, requires a small set of samples. In contrast, in Fig. 15(b), when the same sequence is intentionally removed from the training data, the prediction is less accurate but the larger sample set offsets the inaccuracy.

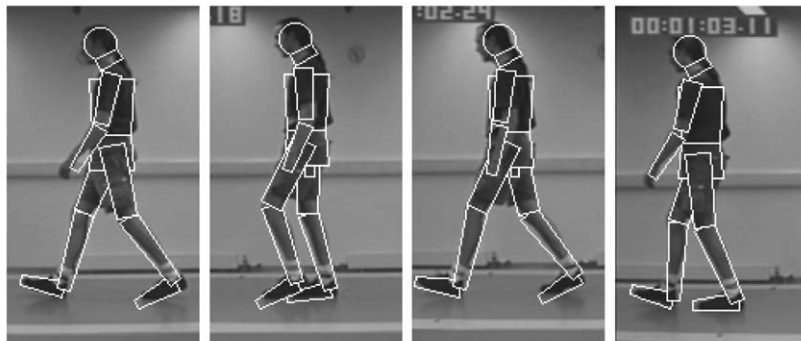
Then a question arises: what a role does the motion model play in tracking? We track a sequence without sampling, so that the motion parameters are estimated entirely from the motion model. The tracking results illustrated in Fig. 16 reveals that, although roughly true, the results are wrong at some body parts in contrast to that in Fig. 13(c). Therefore the motion model does not unduly affect the tracking and the sample set can offset the prediction errors.

Compared with similar previous work, our algorithm requires a much smaller sample set (300 samples for each upper limb and 100 for each lower limb, whereas 500 are used in [13] and 512 in [5]). The effect is due partly to the effectiveness of our dynamic model, and partly to the accurate and robust PEF modeled with a radial term. Further experiments are needed to determine the minimal required number of the sample set in order to make the computational cost as low as possible. However, run-time analysis reveals that most of the time is spent on evaluating the PEF. So a PEF with a lower computational cost will make the overall algorithm more efficient.

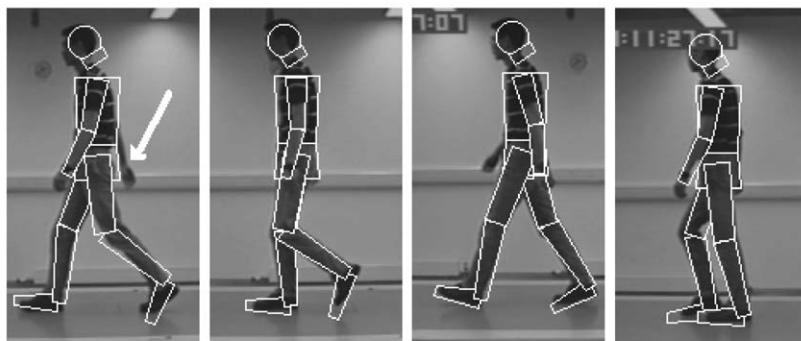
The tracked sequences in Figs. 13 and 14 are all selected from the SOTON and NLPR gait database. In the two databases, the backgrounds are basically clean. So we further test our approach on some additional sequences in more complex real-world outdoor scenes. These sequences were captured on different days and have significant motion blur and changing background due to winds. Two samples of such sequences are shown in Fig. 17. It can be seen that the tracking results are fairly accurate except some errors on the upper limbs superposing on the torso (e.g. frame 32 in Fig. 17(a) and frame 38 in Fig. 17(b)). These errors are mainly caused by the noticeable motion blur where the edges of most of the limbs can hardly be found when the limbs are superposing on other body parts (e.g. frame 32 in Fig. 17(a) and frame 38 in Fig. 17(b) where the arms are pulled to the torso edges). For this specific issue where the arm is naked, the skin color model may be used to segment out the arms. But the more applicable method is to use image restoration to remove the motion blur or reduce the time of exposure. Finally, it is worthwhile to mention the applicability of our motion model. The motion model is learnt from the SOTON gait database where the subjects are mainly male and European, but it is still applicable to the two Chinese girls shown in Fig. 17. However, it often fails when it is applied to recover the poses of their arms off the camera. This means that we should extend the motion model so that to cover more individuals.



(a) Frame 19, 27, 42 and 53 in sequence 1 of subject 6 (sg1)



(b) Frame 19, 29, 35 and 47 in sequence 3 of subject 6 (sg3)



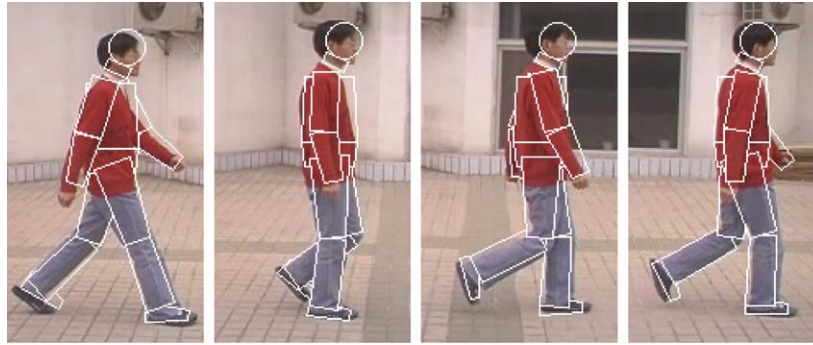
(c) Frame 19, 23, 33 and 43 in sequence 2 of subject 3 (vin2)

Fig. 13. Tracking persons in training data.

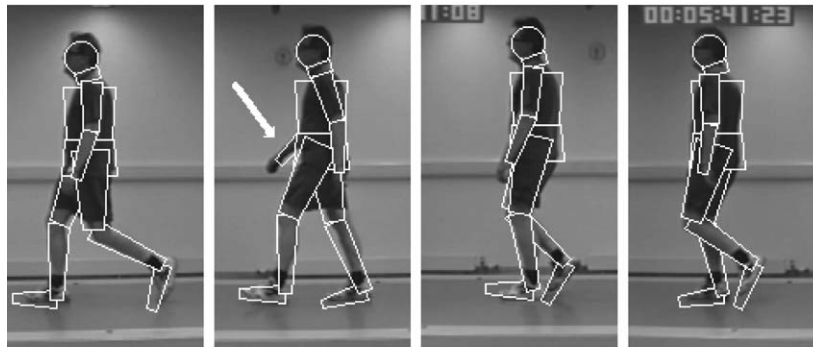
7.3. Gait recognition

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. Fig. 18 shows the temporal curves of shoulder and knee angles. The motion parameters 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 to further illustrate the usefulness of the proposed tracking algorithm (It should be mentioned that a small portion of the dynamic data are acquired using our previous approach [34].). 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,



(a) Frame 18, 24, 37 and 53 in sequence 2 of wl in NLPR gait database



(b) Frame 19, 27, 38 and 53 in sequence 1 of subject 1 (dh1)

Fig. 14. Tracking persons not in training data.

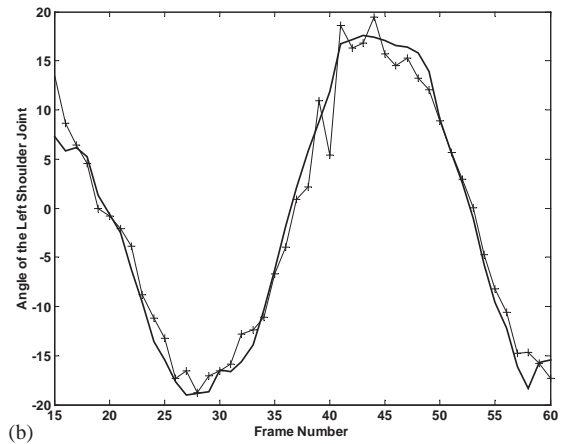
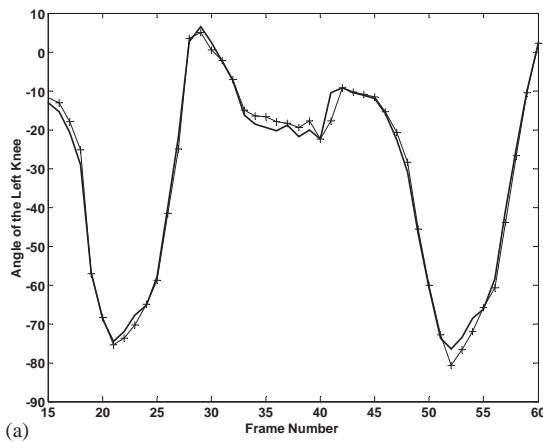


Fig. 15. Estimation results: predicted results are in thin lines with markers and refined results by factored sampling are in bold lines, (a) the angle of the left knee of dh1 that is included in the training data; (b) the angle of the left shoulder joint of the same sequence that is removed from the training data.

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 pro-

posed 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. 19 shows performance of

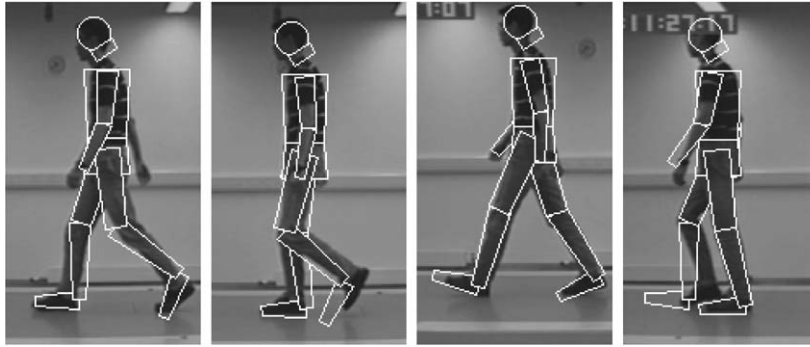
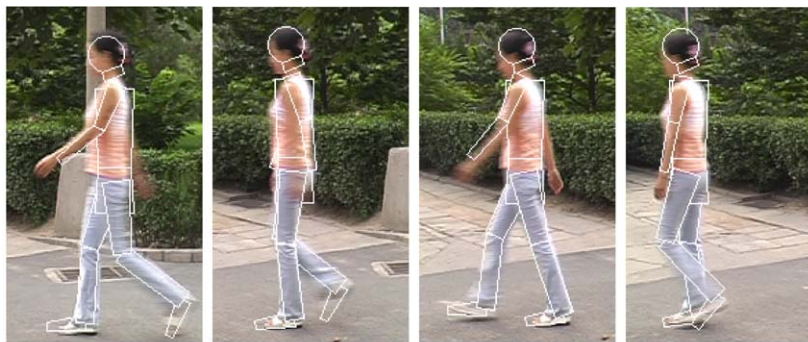
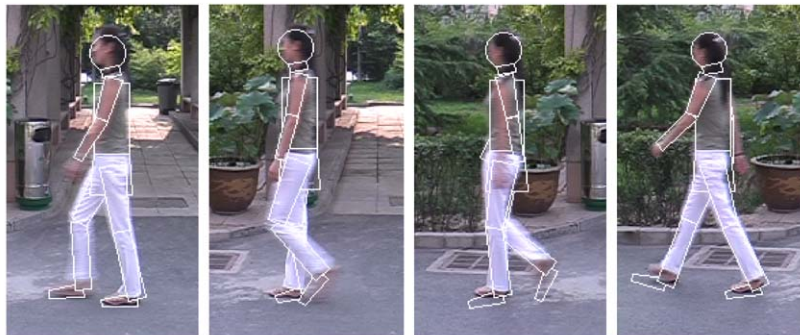


Fig. 16. Analyzing the significance of the motion model. Tracking results of frame 19, 23, 33 and 43 in sequence 2 of subject 3 (vin2) without sampling.



(a) Frame 15, 32, 37 and 47 in sequence 1 of yy.



(b) Frame 15, 25, 38 and 46 in sequence 3 of mh.

Fig. 17. Tracking persons in complex real-world outdoor scenes.

identification and verification using dynamic and static information, respectively. The recognition results using static features are detailed in [33]. 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.

7.4. Synthetic walking

With the posture vectors of the tracking results, we can also synthesize the walking process with parameterized 3D articulated models. The synthetic sequence, which can be inspected from any viewing angle, gives us a more vivid impression of walking movement. Fig. 20 gives a synthetic example of walking from far to near along 45° direction.

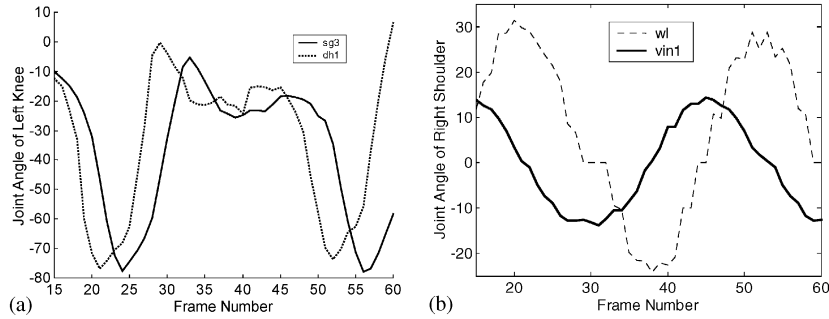


Fig. 18. Analysis of tracking results, (a) temporal curve of joint angle of the left knee; (b) temporal curve of joint angle of the right shoulder.

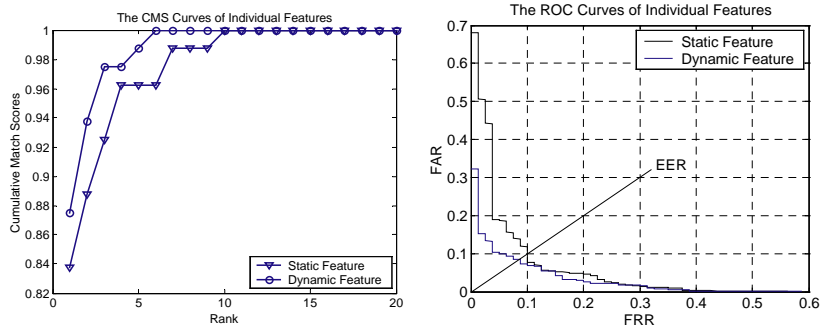


Fig. 19. Gait-based human identification and verification performance.

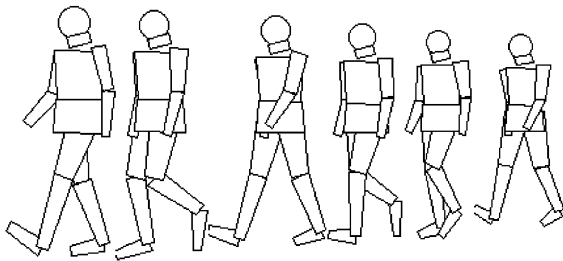


Fig. 20. Synthetic walking corresponding to the sequence 1 of subject 1 (dh1).

The synthesis is a little coarse because of the simplicity of our human body model. With a more complex model, the temporal data of human motions can be used to achieve better results of animation.

8. Conclusion and future work

We have presented a novel algorithm for people tracking in monocular image sequences using the CONDENSATION framework. A motion model was learnt from the semi-automatically acquired training data, and motion constraints were explored by analyzing the dependency of joints. Then both of them were integrated into the dynamic model

to reduce the size of the sample set by concentrating the samples in the areas of the state space containing most information about the posterior. Another contribution is the automatic initialization and the capability of recovery from severe failures, which were achieved using spatio-temporal information. We had also defined a novel pose evaluation function combing both boundary and region information to compute the observation density. Experiments have shown the effectiveness of our method, and the analysis of the results encourages us to do further research on gait recognition using such high-level information as temporal features of joint angles.

Several issues remain to be studied. Although the initial pose is automatically acquired, the stationary parameters of human body model, such as radius and length of each part, are actually measured by hand before tracking. So it is necessary to develop some techniques to estimate the stationary parameters to make our system fully automatic. We are also developing a faster algorithm to compute the PEF that is the most time-consuming in most model-based tracking systems. Additionally, we are extending our motion model and motion constraints to better use the prior information in the CONDENSATION framework. Currently, 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 human motion capturing in unconstrained environments, much work still remains to be done.

Acknowledgements

The authors would like to thank Dr. M. Nixon and Dr. C. Yam from University of Southampton, UK, for their help with the SOTON gait database. This work is supported in part by NSFC (Grant No. 60121302, 69825105 and 60105002) and the National High Tech R& D Program of China (Grant No. 2002AA117010).

References

- [1] D. Gavrilu, The visual analysis of human movement: a survey, *Comput. Vis Image Und.* 73 (1) (1999) 82–98.
- [2] J. Aggarwal, Q. Cai, Human motion analysis: a review, *Comput. Vis. Image Und.* 73 (3) (1999) 428–440.
- [3] J.D. Shutler, M.S. Nixon, C.J. Harris, Statistical gait recognition via temporal moments, in: *Proceedings of the Fourth IEEE Southwest Symposium on Image Analysis and Interpretation*, 2000, pp. 291–295.
- [4] J. Foster, M. Nixon, A.P. Bennett, New area based metrics for gait recognition, in: *Proceedings of the Third International Conference on Audio- and Video-Based Biometric Person Authentication*, 2001, pp. 312–317.
- [5] T. Zhao, T.S. Wang, H.Y. Shum, Learning a highly structured motion model for 3D human tracking, in: *Proceedings of Fifth Asian Conference on Computer Vision*, Melbourne, Australia, 2002.
- [6] H.J. Lee, Z. Chen, Determination of 3D human body posture from a single view, *Comput. Vis. Graph.* 30 (1985) 148–168.
- [7] D. Hogg, Model-based vision: a program to see a walking person, *Image Vision Comput.* 1 (1) (1983) 5–20.
- [8] S. Wachter, H.H. Nagel, Tracking persons in monocular image sequences, *Comput. Vision Image Und.* 74 (3) (1999) 174–192.
- [9] Q. Delamarre, O. Faugeras, 3D articulated models and multi-view tracking with physical forces, *Comput. Vis. Image Und.* 81 (2001) 328–357.
- [10] Q. Delamarre, O. Faugeras, 3D articulated models and multi-view tracking with silhouettes, in: *Proceedings of the Seventh International Conference on Computer Vision*, Kerkyra, Greece, 1999.
- [11] C. Sminchisescu, B. Triggs, Covariance scaled sampling for monocular 3D body tracking, in: *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, Kauai, HI, 2001.
- [12] R. Plankers, P. Fua, Articulated soft objects for video-based body modeling, in: *Proceedings of the Ninth International Conference on Computer Vision*, Vancouver, Canada, 2001.
- [13] H. Sidenbladh, M. Black, David Fleet, Stochastic tracking of 3D human figures using 2D image motion, in: *Proceedings of the European Conference on Computer Vision*, 2000.
- [14] 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.
- [15] C. Bregler, Learning and recognizing human dynamics in video sequences, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, 1997.
- [16] M. Murray, Gait as a total pattern of movement, *Ame. J. Physical Med.* 46 (1967) 290–333.
- [17] M. Isard, A. Blake, Condensation—Conditional density propagation for visual tracking, *Int. J. Comput. Vision* 29 (1) (1998) 5–28.
- [18] M. Isard, A. Blake, A mixed-state condensation tracker with automatic model switching, in: *Proceedings of the International Conference on Computer Vision*, 1998, pp. 107–112.
- [19] E. Ong, S. Gong, A dynamic human model using hybrid 2D-3D representation in hierarchical PCA space, in: *Proceedings of the 10th British Machine Vision Conference*, United Kingdom, 1999.
- [20] J. Hoshino, H. Saito, A match moving technique for merging CG and human video sequences, in: *Proceedings of the ICASSP*, 2001.
- [21] David G. Lowe, Fitting parameterized 3-D models to images, *IEEE T. on Pattern Anal.* 13 (1991) 441–450.
- [22] W. Gilks, S. Richardson, D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, Chapman and Hall, 1996.
- [23] M. Isard, A. Blake, Icondensation: unifying low-level and high-level tracking in a stochastic framework, *Proceedings of the European Conference on Computer Vision 1* (1998) 893–909.
- [24] A. Bottino, A. Laurentini, A silhouette based technique for the reconstruction of human movement, *Comput. Vis. Image Und.* 83 (2001) 79–95.
- [25] C. Sminchisescu, B. Triggs, A robust multiple hypothesis approach to monocular human motion tracking, *Technical Report RR-4208*, INRIA, 2001.
- [26] D. Magee, R. Boyle, Detecting lameness in livestock using resampling condensation and multi-stream cyclic hidden Markov models, in: *Proceedings of the British Machine Vision Conference*, 2000, pp. 332–341.
- [27] T. Moeslund, E. Granum, A survey of computer vision-based human motion capture, *Comput. Vis. Image Und.* 81 (2001) 231–268.
- [28] A. Pentland, Looking at people: sensing for ubiquitous and wearable computing, *IEEE T. Pattern Anal.* 22 (1) (2000) 107–119.
- [29] I.A. Karaulova, P.M. Hall, A.D. Marshall, A hierarchical model of dynamics for tracking people with a single video camera, *British Machine Vision Conference*, (2000) 352–361.
- [30] K. Rohr, Towards model-based recognition of human movements in image sequences, *CVGIP: Image Und.* 59 (1) (1994) 94–115.
- [31] J. B. Hayfron-Acquah, M.S. Nixon, J.N. Carter, Automatic gait recognition by symmetry analysis, in: *Proceedings of the Third International Conference on Audio- and Video-based Biometric Person Authentication*, 2001, pp. 272–277.
- [32] 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) 383–388.
- [33] L. Wang, T. Tan, W. Hu, H. Ning, Automatic gait recognition based on statistical shape analysis, *IEEE T. Image Process* 12 (9) (2003) 1120–1131.
- [34] H. Ning, L. Wang, W. Hu, T. Tan, Model-based tracking of human walking in monocular video sequences, *The 17th IEEE Region 10 Technical Conference on Computers, Communication, Control and Power Engineering*, 2002.

About the Author —HUAZHONG NING received his B.Sc. in the Department of Special Class for Gifted Young from University of Science and Technology of China, with major of Computer Science, in 2000 and M.Sc in pattern recognition and intelligent system from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2003, respectively. He has published more than 8 papers on international journals and conferences. His main research interests include computer vision, video computing, tracking, human computer interaction, image processing, pattern recognition, graphics, etc.

About the Author —TIENIU TAN (M'92-SM'97) received his B.Sc. in Electronic Engineering from Xi'an Jiaotong University, China, in 1984 and M.Sc., DIC and Ph.D. in Electronic Engineering from Imperial College of Science, Technology and Medicine, London, UK, in 1986, 1986, and 1989, respectively. He joined the Computational Vision Group at the Department of Computer Science, The University of Reading, England, in October 1989, where he worked as Research Fellow, Senior Research Fellow and Lecturer. In January 1998, he returned to China to join the National Laboratory of Pattern Recognition, the Institute of Automation of the Chinese Academy of Sciences, Beijing, China. He is currently Professor and Director of the National Laboratory of Pattern Recognition as well as President of the Institute of Automation. He has published widely on image processing, computer vision and pattern recognition. His current research interests include speech and image processing, machine and computer vision, pattern recognition, multimedia, and robotics. He serves as referee for many major national and international journals and conferences. He is an Associate Editor of Pattern Recognition and IEEE Transactions on Pattern Analysis and Machine Intelligence, the Asian Editor of Image and Vision Computing. Dr. Tan is an IEEE Fellow now. He was an elected member of the Executive Committee of the British Machine Vision Association and Society for Pattern Recognition (1996–1997) and is a founding co-chair of the IEEE International Workshop on Visual Surveillance.

About the Author —LIANG WANG received his B.Sc. in Electrical Engineering and M.Sc. in Video Processing & Multimedia Communication from Anhui University, Hefei, China, in 1997 and 2000, and Ph.D in Pattern Recognition & Intelligent System from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2003, respectively. He has published more than 16 papers on major international journals and conferences. His current research interests include computer vision, pattern recognition, digital image processing and analysis, multimedia, visual surveillance, etc.

About the Author —WEIMING HU received his Ph.D. from the Department of Computer Science and Engineering, Zhejiang University, Hangzhou, China. From April 1998 to March 2000, he worked as a Postdoctoral Research Fellow at the Institute of Computer Science and Technology, Founder Research and Design Center, Peking University. From April 2000, he worked at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, as an Associate Professor. His research interests are in visual surveillance and monitoring of dynamic scenes, neural network, 3D computer graphics, physical design of ICs, and map publishing system. He has published more than 20 papers on major national and international journals and conferences.