Fusion of Static and Dynamic Body Biometrics for Gait Recognition

Liang Wang, Huazhong Ning, Tieniu Tan, *Fellow, IEEE*, and Weiming Hu

***Abstract*—Vision-based human identification at a distance has recently gained growing interest from computer vision researchers. This paper describes a human recognition algorithm by combining static and dynamic body biometrics. For each sequence involving a walker, temporal pose changes of the segmented moving silhouettes are represented as an associated sequence of complex vector configurations and are then analyzed using the Procrustes shape analysis method to obtain a compact appearance representation, called static information of body. In addition, a model-based approach is presented under a Condensation framework to track the walker and to further recover joint-angle trajectories of lower limbs, called dynamic information of gait. Both static and dynamic cues obtained from walking video may be independently used for recognition using the nearest exemplar classifier. They are fused on the decision level using different combinations of rules to improve the performance of both identification and verification. Experimental results of a dataset including 20 subjects demonstrate the feasibility of the proposed algorithm.**

***Index Terms*—Biometrics, gait recognition, joint-angle trajectory, Procrustes shape analysis, tracking.**

I. Introduction

IDENTIFYING people automatically and accurately is an important task in a full range of visual surveillance and monitoring applications. In controlled environments such as airports, banks, and car parks, it is desirable to quickly detect threats. Recent events such as the September 11th attack have brought biometrics (especially noncontact human identification at a distance) to the frontline of attention.

A. *Motivation of Gait Recognition*

Gait is a newly emergent biometric feature which offers the ability of identifying people at a distance. Gait can be advantageous in some aspects over other forms of biometric features in the following ways.

1) Gait seems to be unique. That each person seems to have a distinctive way of walking is easily understood from a biomechanics viewpoint [11]. Human walking is a complex action of locomotion involving synchronized integrated movements of body parts, joints, and the interaction among them [11]. It is the distinguishable variations among the properties of body structures, weights of limbs, and actions of different subjects that may provide a unique cue for identity recognition.
2) Gait is unobtrusive. Most biometric features usually require physical touch or proximal sensing, while using gait would avoid such problems since it does not require the user's interaction. Also, gait can be easily extracted from great distances secretly, which naturally advances the acceptance of the users.
3) Gait can be used for recognition at a distance. The established biometric features such as face and fingerprint are limited in such a capability because they usually require sensing the cooperative users at close ranges. However, at a distance, these biometric features are hardly applicable. Fortunately, gait is still visible in this case. So, from the surveillance point of view, gait is a very attractive modality for recognition at a distance.

As stated above, gait has many advantages, especially unobtrusive identification at a distance, making it very attractive. Gait recognition, as a combination of human motion analysis [5] and biometrics, aims essentially to discriminate people by the way they walk. An ongoing research project, the Human Identification at a Distance (Human ID) program[1] [24] sponsored by DARPA, aims to develop a full range of multimodal surveillance technologies for detecting, classifying, and identifying humans from a great distance to enhance protection from terrorist attacks. Its focus is on dynamic face recognition and recognition from body dynamics including gait.

B. *Related Work*

Interest in automatic gait recognition in the computer vision community only began recently, but considerable efforts have already been made [2], [4], [11], [12], [14]–[16], [23], [26]–[28]. These methods can be roughly divided into two major categories, namely model-based methods and motion-based methods.

Model-based approaches [2], [4], [14], [26] usually model the human body structure or motion and extract image features to map them into the model components. For instance, Johnson and Bobick [4] used activity-specific static body parameters for gait recognition without directly analyzing gait dynamics. Yam *et al.* [14] first used running to recognize people as well as walking and explored the relationship between walking and running that was expressed as a mapping based on phase modulation. Cunado *et al.* [26] used thigh joint trajectories as features. The advantages of model-based approaches are that they offer

Manuscript received November 1, 2002; revised August 25, 2003. This work was supported in part by the National Science Foundation of China under Grant 60121302, Grant 60332010, and Grant 60105002, the National 863 High-Tech R&D Program of China under Grant 2002AA117010, and the Natural Science Foundation of Beijing under Grant 4031004. An earlier and shorter version of this paper appeared in [25].

The authors are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, P. R. China (e-mail: lwang@nlpr.ia.ac.cn; hzning@nlpr.ia.ac.cn; tnt@nlpr.ia.ac.cn; wmhu@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TCSVT.2003.821972

[1] Available. [Online]. http://www.darpa.mil/iao/HID.htm

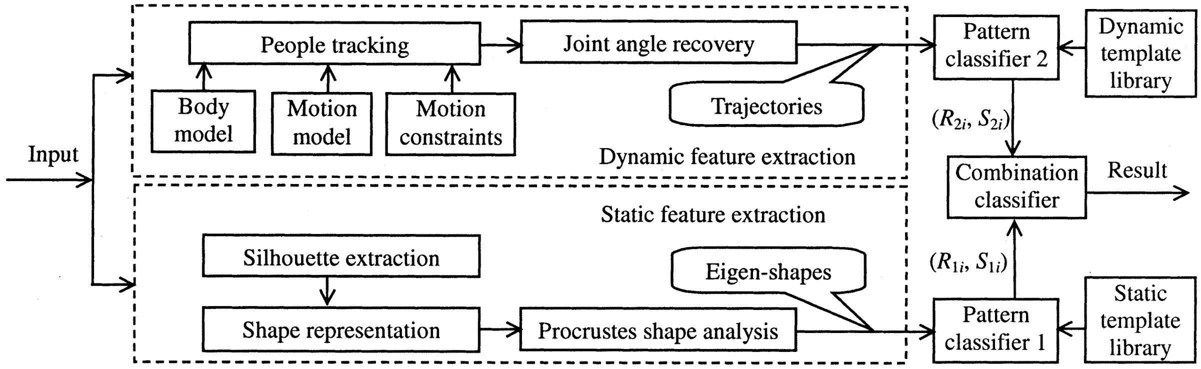


Fig. 1. Overview of the proposed algorithm.

the ability to derive gait signatures directly from model parameters. The disadvantage is that the computational cost is high due to the complex matching and searching.

Motion-based approaches [11], [12], [15], [16], [23], [27], [28] generally characterize the whole motion pattern of the human body by a compact representation regardless of the underlying structure. For example, BenAbdelkader *et al.* [11] proposed an eigengait method using image self-similarity plots, Collins *et al.* [12] established a method based on template matching of body silhouettes in key frames during a walk cycle for human identification, and Phillips *et al.* [15] described a baseline algorithm based on spatial-temporal silhouette correlation for the gait identification problem. In comparison, motion-based approaches are of lower computational complexity and simpler implementation.

These early results further confirm that gait has a rich potential for human identification. Compared with other widely used biometric features such as face and fingerprint, gait recognition is still in its infancy. Vision-based gait recognition will thus offer us an interesting research topic.

C. Overview of the Approach

For obtaining optimal performance, an automatic person identification system should incorporate as many informative cues as available. There are many properties of gait that might serve as recognition features. We categorize them as static features and dynamic features. The former usually reflect geometry-based measurements such as body height and build, while the latter mean joint-angle trajectories of main limbs. Intuitively, recognizing people by gait depends greatly on how the static silhouette shape changes over time. Thus, most previous work on gait recognition mainly adopted low-level information such as silhouette [11], [12], [15], [16]. Due to the difficulties of automatic parameter recovery from video, few methods except [14] and [26] used higher level information, e.g., temporal features of joint angles reflecting the gait dynamics sufficiently. Based on the idea that body biometrics includes both the appearance of the human body and the dynamics of gait motion measured during walking [16], here we attempt to fuse the two completely different sources of information available from the walking video for personal recognition.

The proposed method is schematically shown in Fig. 1. For each image sequence, background subtraction is used to extract moving silhouettes of the walker. Temporal pose changes of these silhouettes are represented as an associated sequence of complex vector configurations in a common coordinate and are then analyzed using the Procrustes shape analysis method to obtain an eigenshape for reflecting the body appearance, i.e., static information. Also, a model-based approach under a Condensation framework together with human body model, motion model, and constraints is presented to track the walker in image sequences. From the tracking results, we can easily calculate joint-angle trajectories of the main lower limbs, i.e., dynamics of gait. Both static and dynamic information may be independently used for recognition using the nearest exemplar pattern classifier. They are also combined effectively on a decision level to improve recognition performance. This method is in essence a combination of model-based and motion-based approaches. It not only analyzes the spatio-temporal motion pattern of gait dynamics but also derives a compact statistical appearance description of gait as a continuum. Thus, it implicitly captures both structural (appearances) and transitional (dynamics) characteristics of gait.

The remainder of this paper is organized as follows. Sections II and III describe static and dynamic feature extraction, respectively. Pattern classifiers and combination rules are presented in Section IV. Section V provides experimental results prior to conclusions in Section VI.

II. Static Feature Extraction

A. Silhouette Extraction and Representation

To segment the walking figure from the background image, a change-detection procedure [23] is adopted to extract a single-connectivity moving region in each frame. An important cue in determining underlying motion of a walking figure is his or her temporal changes of silhouette shape. For the sake of reducing redundancy, here we only need to analyze spatial contours. The extraction and representation process of the silhouette is illustrated in Fig. 2. The silhouette's boundary can be obtained using a border-following algorithm based on connectivity. Then, we may compute its shape centroid (x_c, y_c). Let the centroid be the origin of a two-dimensional (2-D) shape space. We can unwrap the boundary as a set of pixel points (x_i, y_i) along the outer contour counterclockwise in a complex coordinate. That is, each shape can be described as a vector consisting of complex num-

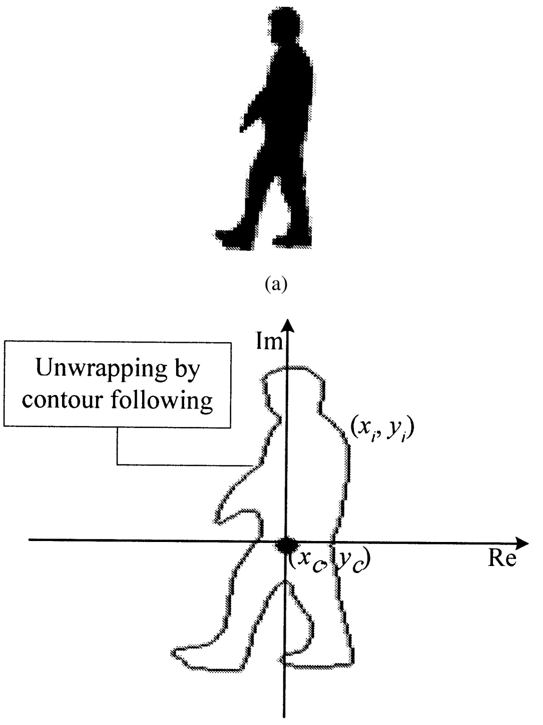


Fig. 2. Illustration of silhouette shape representation. (a) Moving sillhouette. (b) Boundary unwrapping.

bers with N_b boundary elements $z = [z_1, z_2, \ldots, z_i, \ldots z_{Nb}]^T$, where $z_i = x_i + j^* y_i$. Each gait sequence will be accordingly transformed into a sequence of such 2-D shape configurations.

B. Procrustes Shape Analysis

We need one method that allows us to compare a set of static pose shapes in gait pattern and is robust to the changes of position, scale, and slight rotation. A mathematically elegant way for aligning point sets is the Procrustes shape analysis [1].

The following gives a summary almost taken from [1] for self-containedness. Procrustes shape analysis is intended to cope with 2-D shapes. A shape in 2-D space can be described by a vector of k complex numbers, $z = [z_1, z_2, \ldots, z_k]^T$, called a configuration. For two shapes, z_1 and z_2, if their configurations are equal through a combination of translation, scaling, and rotation [1]

$$\begin{cases} \mathbf{z}_1 = \alpha \mathbf{1}_k + \beta \mathbf{z}_2, \alpha, \beta \in C \\ \beta = |\beta| e^{j\angle\beta} \end{cases} \tag{1}$$

where $\alpha\mathbf{1_k}$ translates $\boldsymbol{z_2}$, and $|\beta|$ and $\angle\beta$ scale and rotate $\boldsymbol{z_2}$, we may consider that they are the same shape. To center shapes, the centered configuration is defined as $\boldsymbol{u} = [u_1, u_2, \ldots, u_k]^T$, $u_i = z_i - \bar{z}$, $\bar{z} = \sum_{i=1}^{k} z_i / k$. The full Procrustes distance between two configurations $\boldsymbol{u_1}$ and $\boldsymbol{u_2}$ can be defined by

$$d_F(\boldsymbol{u}_1, \boldsymbol{u}_2) = 1 - \frac{|\boldsymbol{u}_1^* \boldsymbol{u}_2|^2}{\|\boldsymbol{u}_1\|^2 \, \|\boldsymbol{u}_2\|^2} \tag{2}$$

where the superscript * represents the complex conjugation transpose. Given a set of n shapes, we can find their mean $\boldsymbol{u}$ that minimizes the objective function [1]

$$\min_{\alpha_j, \beta_j} \sum_{j=1}^{n} \|\boldsymbol{u} - \alpha_j \mathbf{1_k} - \beta_j \boldsymbol{u}_j\|^2. \tag{3}$$

To find $\boldsymbol{u}$, we compute the following matrix:

$$\boldsymbol{S_u} = \sum_{i=1}^{n} \frac{(\boldsymbol{u}_i \boldsymbol{u}_i^*)}{(\boldsymbol{u}_i^* \boldsymbol{u}_i)}. \tag{4}$$

The Procrustes mean shape $\hat{\boldsymbol{u}}$ is the dominant eigenvector of $\boldsymbol{S}_u$, i.e., the eigenvector that corresponds to the greatest eigenvalue of $\boldsymbol{S}_u$ [1].

C. Static Signature Acquisition

Our approach uses these single shape representations from a gait sequence to find their mean shape as a static signature that can implicitly represent the appearance of the body structure. The following summarizes the major steps in determining the Procrustes mean shape for a sequence of shapes from n frames in a gait sequence.

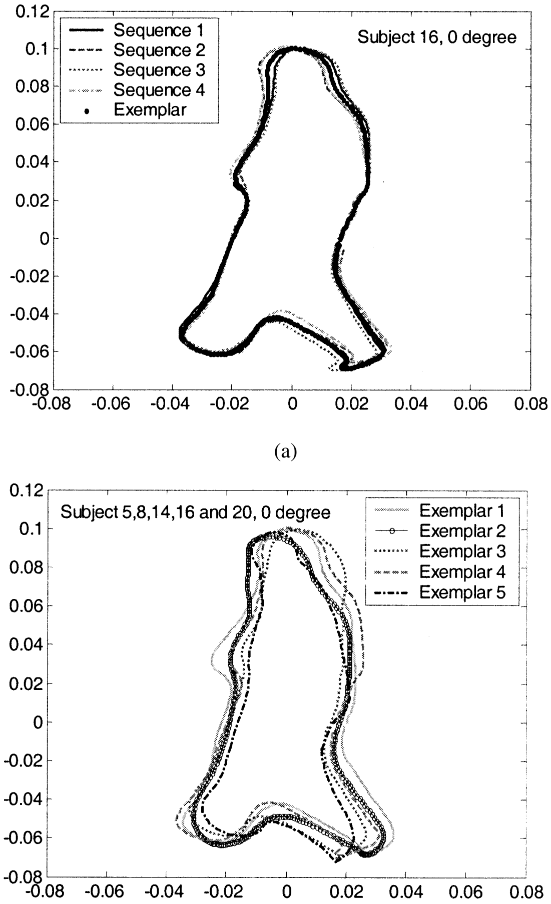
1) Select a set of k points from the boundary to represent a 2-D shape as a vector configuration $\boldsymbol{z}_j$. Here, we tackle the point correspondence problem through interpolation of boundary pixels so that the set is the same for each image.
2) Set the centered configuration. When we represent the silhouette shape, we use the shape centroid as the origin of the 2-D shape space to move all shapes to a common center, which can handle translational invariance. So, we can directly set $\boldsymbol{u}_j = \boldsymbol{z}_j$, $j = 1, 2, \ldots, n$.
3) Compute the matrix $\boldsymbol{S_u}$ using (4). Then, compute the eigenvalues and the associated eigenvectors of $\boldsymbol{S_u}$.
4) Set the Procrustes mean shape $\hat{\boldsymbol{u}}$ as the eigenvector that corresponds to the maximum eigenvalue, and this mean shape is used as the static signature.

For multiple mean shapes from multiple sequences of the same subject, we may acquire an exemplar by averaging them as a static template for that class so as to avoid selecting a random reference sample. Fig. 3(a) shows plots of mean shapes of four sequences of a subject and their exemplar, and Fig. 3(b) shows plots of multiple exemplars from different subjects. From Fig. 3, we can see that the intrasubject changes in eigenshapes are very small, while the intersubject changes are very significant. Such result implies that the mean shapes have considerable discriminating power. More details on static feature extraction may be found in [23].

III. Dynamic Feature Extraction

For extracting dynamic features of gait motion, we present a new model-based approach to tracking the walker under the Condensation framework [9]. Here, we briefly review some previous work on human body models, motion models, and search strategies in order to put our work in better context.

The geometric structure of human body can usually be represented as a stick figure, 2-D contour, or volumetric model such as cylinder [7], truncated cone [8], [22], and super-quadrics



(a)

(b)

Fig. 3. Plots of: (a) mean shapes and (b) the exemplars.

[10]. Motion models of limbs and joints [6] are widely used in tracking. They serve as prior knowledge to predict motion parameters, to interpret human dynamics, or to constrain the estimation of low-level image measurements. The representation of a human body model in each frame is equivalent to a state vector that indicates the current pose of the tracked person. Pose estimation in a high-dimensional configuration space is intrinsically difficult, so search strategies are often carefully designed to reduce the solution space. Generally, four main categories of search strategies exist: kinematics, Taylor models, Kalman filtering, and stochastic sampling [6]–[10], [21].

A. Human Body Model

Similar to [7], the human body model in this work is composed of 14 rigid body parts, including upper and lower torso, neck, two upper arms, two lower arms, two thighs, two lower legs, two feet, and a head, each of which is represented by a truncated cone except for the head, which is represented by a sphere. Each is connected to another at the joints, the angles of which are represented as Euler angles. The above human body model in its general form has 48 degrees of freedom (DOFs). Under the assumption that walking people are usually captured laterally to obtain more apparent motion cues for gait recognition, the state space can be reduced to a twelve-dimensional (12-D) vector $P = \{x, y, \theta_1, \theta_2, \ldots, \theta_{10}\}$, where (x, y) is the global position of human body and θ_i, $i = 1 \sim 10$, is the joint angles of shoulders, elbows, hips, knees, and ankles.

B. Learning Motion Model and Constraints

As a highly constrained activity, human walking patterns are symmetric, periodical, and of little variation in a wide range of people. Thus, it is easy to learn a compact motion model for human gait. The learning process proceeds in three steps. First, the walking cycles in each training example must be rescaled to the same length and aligned to the same phase. Second, the walking cycles are segmented out from the normalized training examples and represented as W_j with $j = 1 \ldots n$. Lastly, motion model is described by Gaussian functions $G_{k,t}(u_{k,t}, \sigma_{k,t}^2)$ empirically for each joint $k\,(k = 1 \ldots 10)$ at any phase $t\,(t = 1 \ldots T)$ in the walking cycle. The learning and representation of our motion model are compact, but they show great effectiveness in estimation of the prior distribution of the initial pose and prediction of the new pose.

We also derive motion constraints from the training data by further exploring the dependency of the neighboring joints: shoulder and elbow, thigh and knee, and knee and ankle. We assume that the lower arm is driven by the upper arm, and the elbow joint is accordingly determined by the shoulder joint. Therefore, 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. The motion constraints for the knee and ankle joints are learned in the same way. We also derive intervals of the valid value for each motion parameter by specifying its maximal and minimal values. More details on learning motion model and constraints can be found in [22].

C. Tracking

Tracking is equivalent to relating the image data to the pose vector. 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.

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 [9] that uses learned dynamical model, together with visual observations, to propagate the random sample set. The rule of the state density propagation over time is [9]

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

where x_t are the motion parameters at time t, $Z_t = (z_1, z_2, \ldots, z_t)$ is the image sequence up to 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 at the previous time step $p(x_{t-1}|Z_{t-1})$ and three other components: the prior distribution $p(x_o)$, i.e., 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

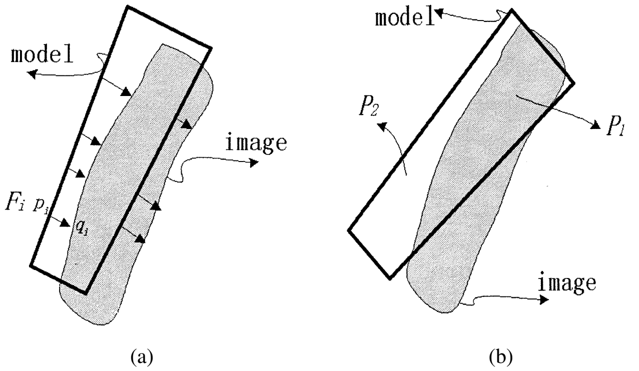


Fig. 4. Measuring the: (a) boundary and (b) region matching errors.

observation density $p(z_t|x_t)$ computed from a pose evaluation function (PEF).

Initialization is concerned with the initial pose of a subject in capturing human motion. Unlike other related work, we use spatio-temporal information of the first N frames to automatically accomplish this process. Due to space limitation, only the dynamic model and the PEF are detailed as follows. Further details may be found in [22].

The dynamic model usually needs to be designed carefully to improve the efficiency of factored sampling. Here, the motion model is integrated into the dynamic model. With the assumption that the Gaussian distributions at different phases in the motion model are independent, at time instant t the ith motion parameter $\theta_{i,t}$ satisfies the dynamic model

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

where G is a Gaussian distribution, $\alpha + \beta + \gamma = 1$ 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. This dynamic model is generally sufficient for all motion parameters, but motion constraints can further concentrate the samples for motion parameters of elbow, knee, and ankle joints. For instance, after the shoulder joint $\theta_{s,t}$ is sampled, sample positions generated from $p(\theta_{e,t}|\theta_{s,t})$ for the elbow joint $\theta_{e,t}$ also contain much useful information. Therefore, a mixed-state Condensation [21] can be included in the factored sampling scheme with a probability q to generate samples from the dynamic model and with a probability $1-q$ to generate samples from the conditional distribution $p(\theta_{e,t}|\theta_{s,t})$, i.e., $\theta_{e,t}$ satisfies

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

where α, β, γ, and λ are defined as given above.

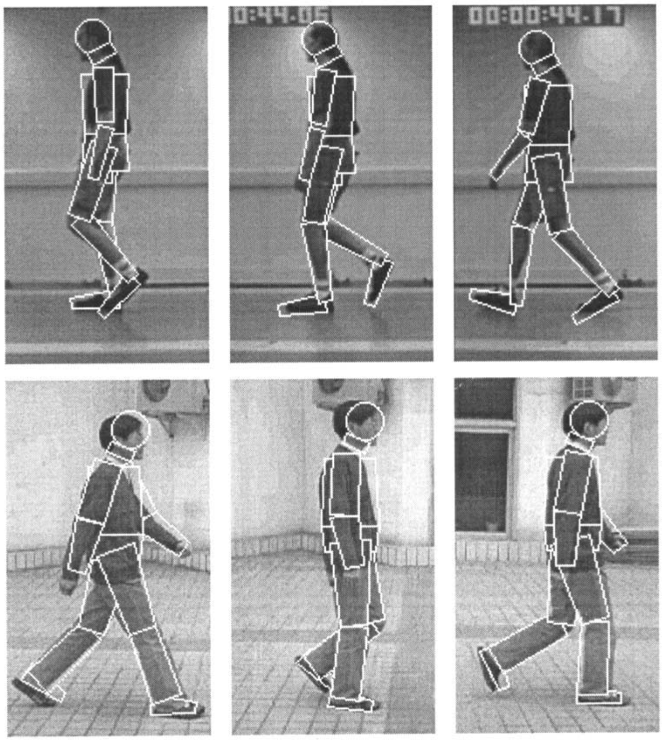


Fig. 5. Part of the tracking results.

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. In general, boundary information improves the localization, whereas region information stabilizes the tracking. Therefore, we combine them in the PEF by computing the boundary and region matching errors simultaneously to achieve both accuracy and robustness.

Fig. 4 shows the procedure of computing the boundary and region matching errors. The boundary matching error E_b is the average distance from the model projection to the boundary of the image that is similar to the Chamfer distance [8]. In computing the region matching error E_r, the region of the projected human model that is fitted into the image data is divided into two parts: P_1 is the region overlapping with the image data and P_2 stands for the rest. Then the matching error is defined by $E_r = |P_2|/(|P_1| + |P_2|)$, where $|P_i|$, $(i = 1, 2)$ is the area, i.e., the number of pixels inside the corresponding region. E_b and E_r are combined into the PEF that is modeled in terms of a radial term $\rho_i(s, \sigma) = ve^{-s/\sigma^2}$ [10] as follows:

$$\mathrm{PEF}(P) = ve^{-(\alpha \times E_b + (1-\alpha) \times E_r)/\sigma^2} \tag{6}$$

where α is a scalar to adjust the weights of E_b and E_r and P is the body pose. Here the tracking results of two sequences are shown in Fig. 5. Due to space constraints, only the human areas clipped from the original images are given here.

D. Dynamic Signature Acquisition

The tracking results enables us to measure joint-angle trajectories. Fig. 6 shows the temporal changes of the angles of four joints: left and right hips, left and right knees from a walking instance, where the smoothed curves are the results after median

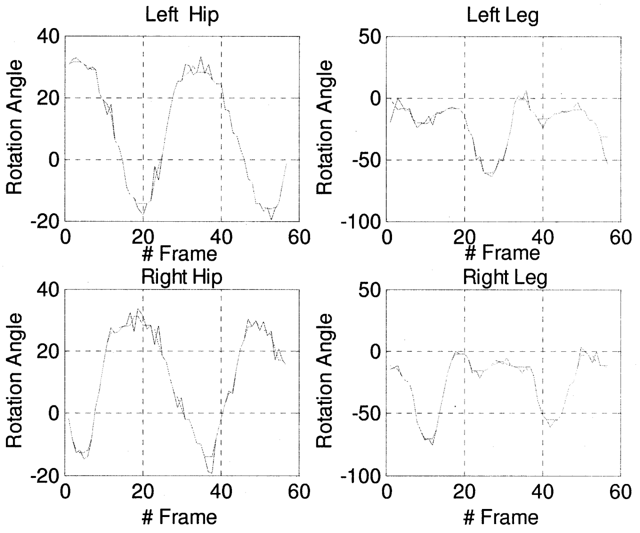


Fig. 6. Joint-angle trajectories of the lower limbs.

filtering. It is the variation in these joint signals that we wish to consider as dynamic information of body biometrics, i.e., gait dynamics, for recognition.

Differences in body structure and dynamics naturally cause joint-angle trajectories to vary in both magnitude and time. So here we normalize them similar to [2] for recognition. Here, we select only one walking cycle from each sequence. Variance normalization by subtracting the mean of each signal and then by dividing by the estimated standard deviation is performed to reduce the effect of noise. Dynamic time warping (DTW) is then applied to temporally align the signals to a fixed reference phase. Fig. 7 shows the results of time-normalized signals of thigh rotation, from which we find that there are little variations among sequences from the same subject, whereas there are apparent variations among different subjects. We choose four normalized signals from the left and right hips and knees to form a dynamic feature vector. Similarly, we also use multiple vectors from the same subject to obtain the exemplar by averaging them, which is regarded as a dynamic template for that class.

IV. Classifiers and Fusion Rules

Gait recognition is a traditional pattern classification problem which can be solved by measuring similarities among gait sequences. Here we try the nearest neighbor classifier with class exemplar (*ENN*). It classifies a sequence as the class of its nearest-neighbor exemplar. There is no doubt that a more sophisticated classifier could be employed, but the interest here is to evaluate the genuine discriminatory ability of the extracted features.

To measure similarity, we make use of the Procrustes mean shape distance defined in (2) for static features and the Euclidean distance for dynamic features, respectively. The smaller the distance measures are, the more similar the two gaits are.

The main reasons for combining classifiers are efficiency and accuracy. A variety of fusion approaches for biometric recognition are available, a few of which are mentioned here [17]–[20]. For example, Hong and Jain [18] integrated face and fingerprint for personal recognition, and a theoretical framework was developed in [19] for combining independent classifiers.

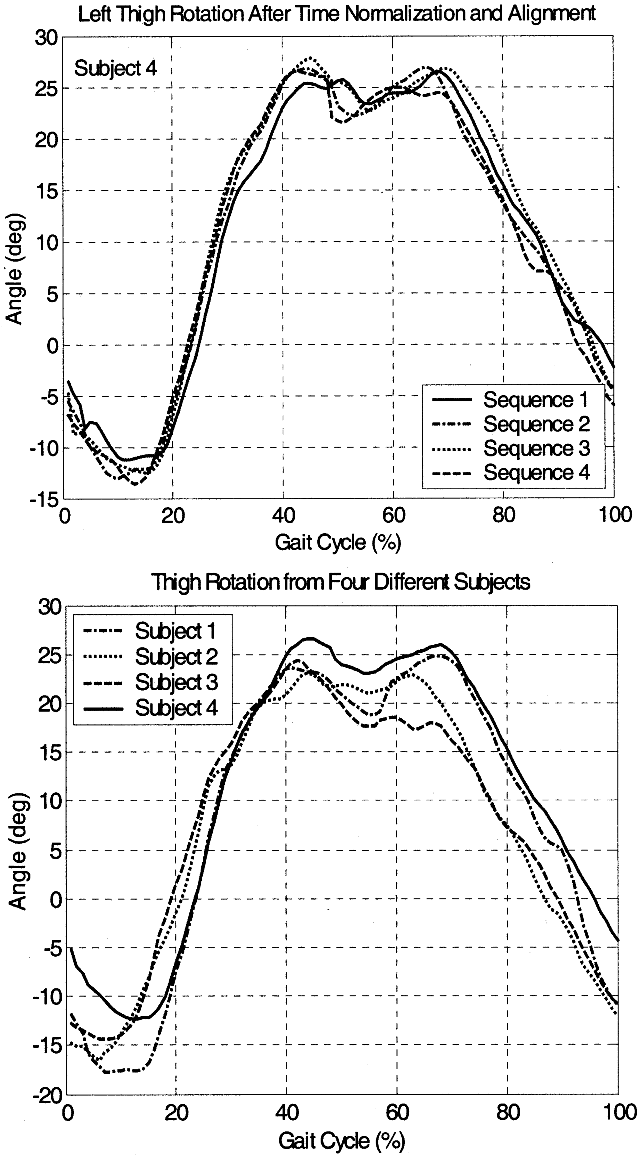


Fig. 7. Time-normalized signals of joint angles.

Having obtained the score for each modality given the observations, one generally cannot directly combine these scores in a statistically meaningful way because these scores are usually not direct estimates of the posterior, but rather measures of the distance between the test examples and the reference example [3]. These scores, with quite different ranges and distributions, must be transformed to be comparable before fusion (the logistic function $e^{(\alpha+\beta x)}/(1+e^{(\alpha+\beta x)})$ is used in this paper).

In this paper, we investigate the following approaches to classifier combination. First, the rank-summation-based and score-summation-based approaches described in [20] are used. The following gives their introductions, which are almost taken from [20]. Rank-based strategies are a generalization of simple voting methods. It is to compute the sum of the rank for every class in the combination set. The class with the lowest rank sum will be the first choice. Let $r(n, R_i)$ be the rank of the

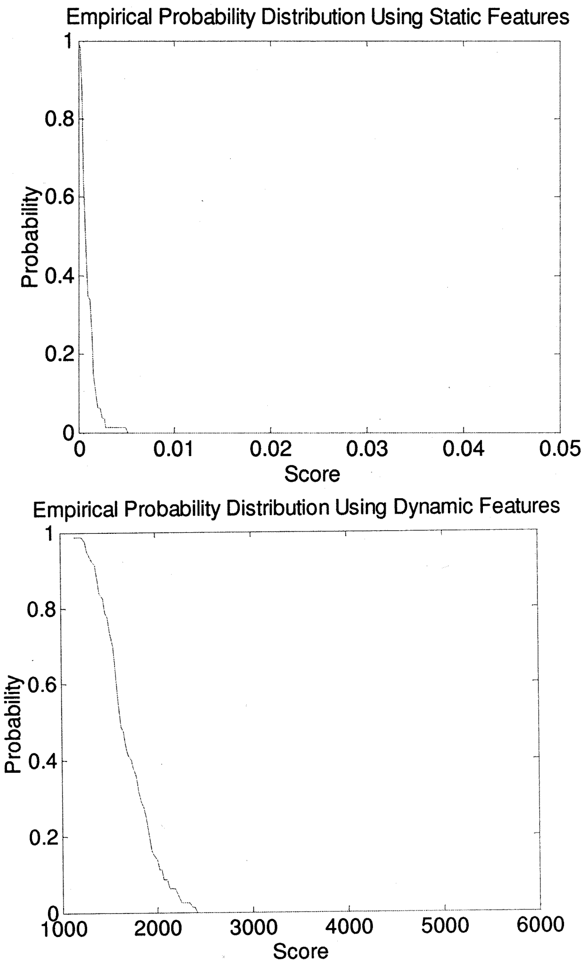


Fig. 8. Modeling the probability distributions of scores.

class with name n in the ranking R_i; this rule is defined as $\arg\min_n \left(n_k, \sum_{j=1}^{R} r(n_k, R_j)\right)$ [20]. If the score functions are directly comparable, the simplest way to combine classifiers using the score is to compute the sum of the score functions. Let $s(n, S_i)$ be the score of the class with name n in the S_i; this rule is defined as $\arg\min_n(n_k, \sum_{j=1}^{R} s(n_k, S_j))$ [20], i.e., the class with the lowest score sum will be the final choice.

Following the theoretical framework presented in [19], we also compare the max, min, mean, and product rules for the combination classifier. Let the input feature to the jth classifier $(j = 1, \ldots, R)$ be $\boldsymbol{x}_j$ and the winning label be l. These rules are given as follows according to [3]:

1) The product rule: $l = \arg\max_k \prod_{j=1}^{R} p(\omega_k/x_j)$.
2) The mean rule (sum): $l = \arg\max_k \sum_{j=1}^{R} p(\omega_k/x_j)$.
3) The max rule: $l = \arg\max_k \max_j p(\omega_k/x_j)$.
4) The min rule: $l = \arg\max_k \min_j p(\omega_k/x_j)$.

In order to justify the above rules statistically, a monotonic transformation function over scores S needs to be applied to reflect the posterior probability. We used the similar approach proposed in [3]. That is, we may estimate a probability distribution over the scores assigned to the correct labels by a mapping function T from scores to the empirical distribution and treat $T(S)$ as the estimate of the posterior (see Fig. 8).

V. Experiments

A. Data Acquisition

We collected 80 sequences from 20 different subjects and four sequences per subject for our experiments. Each sequence includes a walking figure, and the walker moves laterally with respect to the image plane at normal cadence in the field of view without occlusion. All image sequences are captured by a stationary digital camera at a rate of 25 frames per second.

B. Experimental Results

For each image sequence, we first extract static features in the manner described in Section II. In addition, we perform the model-based tracking and recover dynamic features in the manner described in Section III. It should be noted that certain scenarios, such as self-occlusion of body parts, shadow under the feet, and the arm and the torso having the same color, and low quality of the image sequences bring challenges to our tracking method. For a small portion of failed tracking images, we manually obtain the motion parameters, as the focus of this paper is not on tracking per se but on gait recognition using the tracking data as dynamic features.

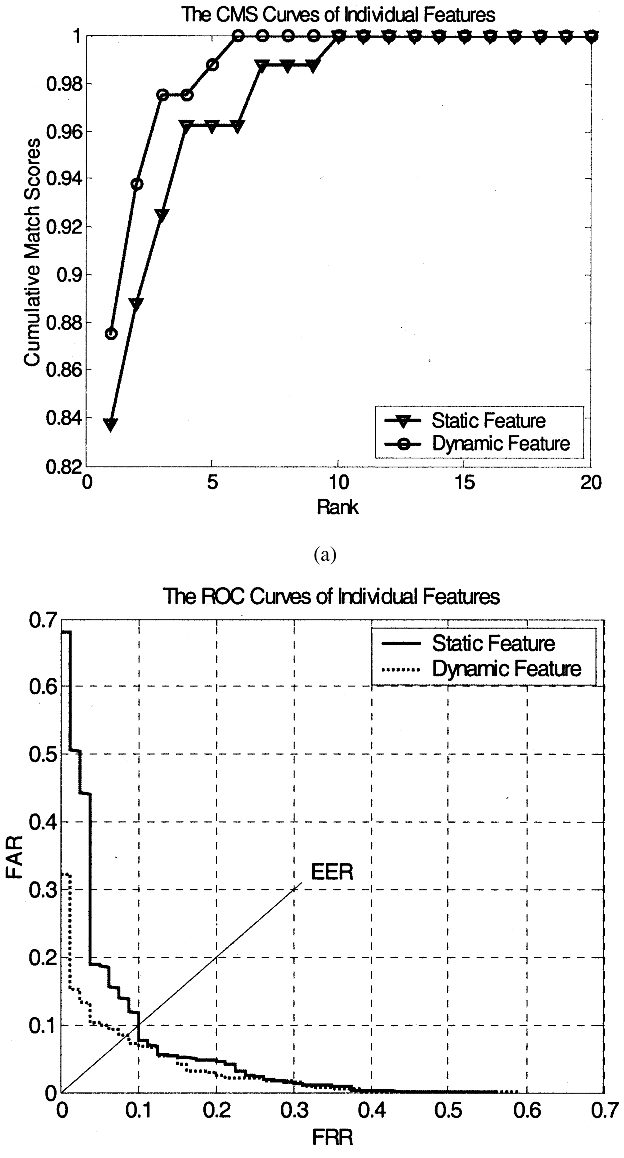
Due to the small number of examples, we hope to compute an unbiased estimate of the true recognition rate using a leave-one-out cross-validation method. That is, we first leave one example out, train on the remainder, and then classify or verify the omitted element according to its differences with respect to the remaining examples.

First, we use static and dynamic features separately for recognition. In the identification mode, the classifier determines to which class a given measurement belongs. One useful measure of classification performance is cumulative match characteristics (CMC) [13] which is first introduced in the FETET protocol for the evaluations of face recognition algorithms. It indicates the probability that the correct match is included in the top n matches. Here, we use it to report the results of identification. For completeness, we also use the receiver operating characteristic (ROC) curves to report verification results. In the verification mode, the classifier is asked to verify whether a new measurement really belongs to certain claimed class. ROC curves give plots of various pairs of false acceptance rate (FAR) and false rejection rate (FRR) under different decision threshold values for the acceptance. Fig. 9(a) and (b) respectively shows performance of identification (for ranks up to 20) and verification using a single modality. It should be mentioned that the correct classification rate (CRR) is equivalent to $p_{(1)}$ (i.e., $\mathrm{rank} = 1$).

Based on the combination rules described in Section IV, we examine the results after fusing both static features and dynamic features. Fig. 10(a) and (b) shows the results of identification and verification using rank-summation-based and score-summation-based combination rules, respectively, and Fig. 11(a) and (b) gives the fusion results using the product, sum, max, and min combination rules, respectively. For comparison, we also plot the results using a single modality in Figs. 10 and 11.

C. Analysis and Discussions

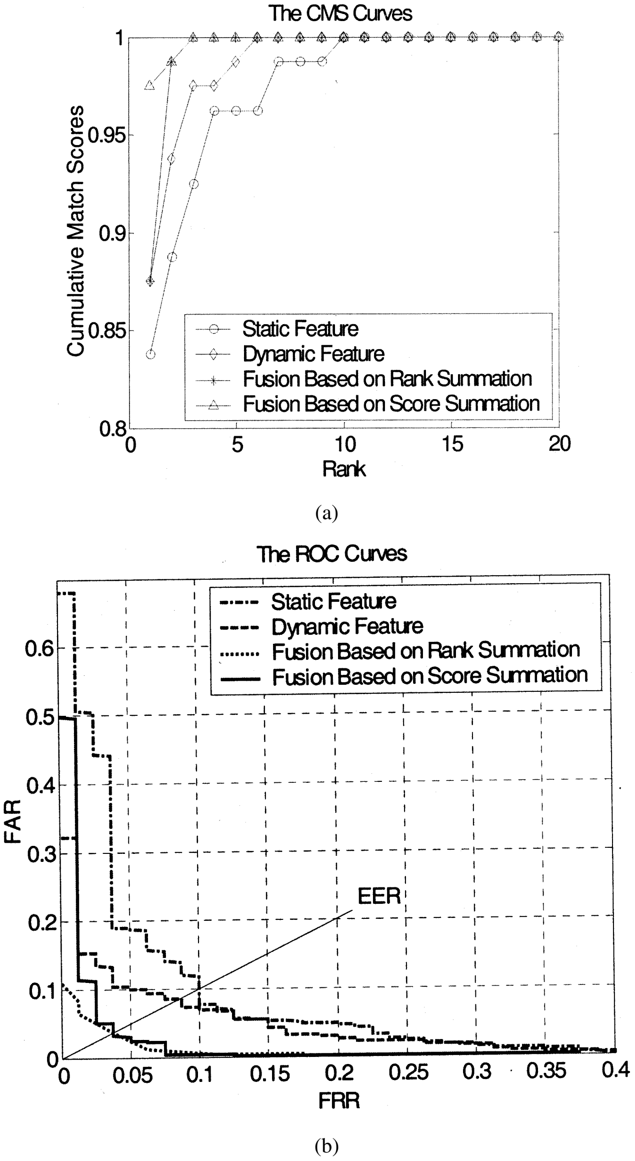
From Fig. 9, we can see that there is indeed identity information in both the static and dynamic features derived from the walking video that can be explored for the recognition task.



(a)

(b)

Fig. 9. Results using a single modality. (a) Identification. (b) Verification.



(a)

(b)

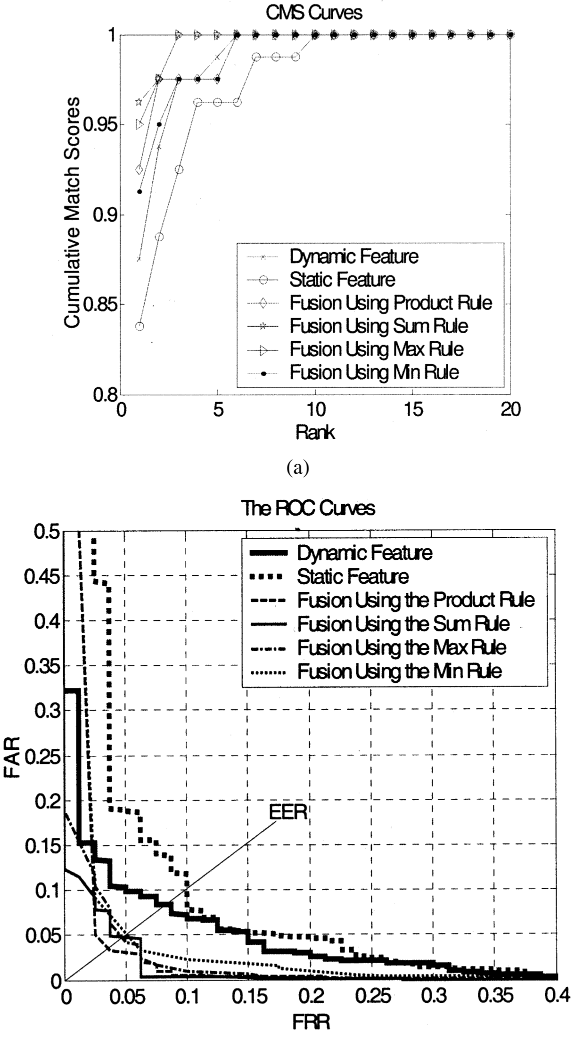
Fig. 10. Results using rank- and score-summation-based rules. (a) Identification. (b) Verification.

The results using dynamic information are somewhat better than those using static information. This is likely due to the fact that the dynamics reflect more essential information of gait motion. As we know, Tanawongsuwan and Bobick [2] also used dynamical features (joint trajectories) for gait recognition, but their work is different from ours. They used a motion capture system to obtain motion data, while in our work the motion parameters were recovered automatically using visual techniques. Also, they achieved a recognition rsate of 73% on a database including 18 subjects while our recognition rate is 87.5% for a database of 20 subjects.

Figs. 10 and 11 demonstrate the improved performance of both identification and verification for the fusion step than that using any single modality. A summary of CCRs and equal error rates (EERs) is given in Table I for clarity. Another observation from the comparative results is that the score-summation-based rule outperforms other combinations schemes as a whole. Of the last four statistical combination rules, the sum rule is the best for identification, which has also been shown in [19] using the sensitivity analysis to demonstrate that the sum rule is the most resilient to estimation errors. However, the product rule is best for verification. It is believed that there will be better results if there are sufficient data to model the probability distributions of scores for the two pattern classifiers more precisely. In all, these studies highlight the importance of a careful choice of the whole combination strategy.

Although the results are very encouraging, more experiments on a larger and more realistic database still need to be further investigated in future work in order to be more conclusive. Accordingly, much remains to be done, which we outline as follows:

1) Establishing a larger and more realistic database. Unlike face recognition, now gait recognition lacks of a common evaluation dataset. The researchers often established their own gait databases and then reported a recognition rate. Directly using these reported rates for comparison seems to mean nothing. We have compared some recent



(a)

(b)

Fig. 11. Results using the product, sum, max, and min combination rules. (a) Identification. (b) Verification.

TABLE I
SUMMARY OF CCRS AND EERS

	CCR (rank=1)	CCR (rank=3)	EER
Static features	83.75%	92.50%	10.0%
Dynamic features	87.50%	97.50%	8.42%
Rank-summation	87.50%	100%	3.75%
Score-summation	97.50%	100%	3.75%
Product	92.50%	97.50%	3.54%
Sum	96.25%	100%	5.00%
Max	95.00%	100%	4.70%
Min	91.25%	97.50%	5.00%

algorithms using static features on our dataset [23]. But this comparison should be considered with reservation. Therefore, we have a strong will that a common dataset will be established so that everyone can make a reasonable comparison with other work.

2) Developing more robust segmentation algorithms and to improve three-dimensional tracking, which is very critical to accurately and automatically extract gait features.
3) Designing more sophisticated classifiers and combination rules.
4) Using a dynamic silhouette description as in [25] in future work to obtain a better description of spatio-temporal silhouette changes in a gait pattern than using static silhouette description here.
5) Further analyzing the correlation of two types of features. As a general rule, the higher the correlation, the lower the recognition rate after fusion. Our two features respectively reveal the two categories of parameters of body biometrics. Since the static parameters of the body are basically un-correlated with the dynamic ones, the silhouette features should be un-correlated with the trajectories to some extent. Also, probably this is the main reason why the fusion performs well in our experiments. Nevertheless, thorough and deep analysis of the correlation is still needed in future work.

VI. CONCLUSION

We have proposed an efficient algorithm based on the fusion of static and dynamic body biometrics for personal recognition. A statistical approach based on the Procrustes shape analysis method is used to obtain a compact representation of the appearance of body shape from the spatio-temporal pattern of the walking action. A model-based approach is employed to track the walker in monocular sequences and to recover joint-angle trajectories of lower limbs that reflect the dynamics of gait motion. Both static and dynamic cues of body biometrics may be independently used for recognition. Also, they have been effectively combined on the decision level for improving performance. Experimental results have demonstrated the feasibility of the proposed algorithm.

ACKNOWLEDGMENT

The authors would like to express their thanks to Dr. Nixon (University of Southampton, U.K.) for providing data and to the anonymous reviewers for their constructive comments.

REFERENCES

[1] J. Boyd, "Video phase-locked loops in gait recognition," in *Proc. Int. Conf. Computer Vision*, vol. I, 2001, pp. 696–703.
[2] R. Tanawongsuwan and A. Bobick, "Gait recognition from time-normalized joint-angle trajectories in the walking plane," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2001.
[3] G. Shakhnarovich, L. Lee, and T. Darrell, "On probabilistic combination of face and gait cues for identification," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, 2002, pp. 176–181.
[4] A. Bobick and A. Johnson, "Gait recognition using static activity-specific parameters," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2001.
[5] L. Wang, W. Hu, and T. Tan, "Recent developments in human motion analysis," *Patt. Recognit.*, vol. 36, no. 3, pp. 585–601, 2003.
[6] T. Zhao, T. Wang, and H. Shum, "Learning a highly structured motion model for 3D human tracking," in *Proc. Asian Conf. Computer Vision*, vol. I, 2002, pp. 144–149.
[7] S. Wachter and H. Nagel, "Tracking persons in monocular image sequences," *Comput. Vis. Image Understanding*, vol. 74, no. 3, pp. 174–192, 1999.

[8] Q. Delamarre and O. Faugeras, "3D articulated models and multi-view tracking with physical forces," *Comput. Vis. Image Understanding*, vol. 81, pp. 328–357, 2001.

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

[10] C. Sminchisescu and B. Triggs, "Covariance scaled sampling for monocular 3D body tracking," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2001.

[11] C. BenAbdelkader, R. Culter, H. Nanda, and L. Davis, "EigenGait: motion-based recognition of people using image self-similarity," in *Proc. Int. Conf. Audio- and Video-Based Person Authentication*, 2001, pp. 284–294.

[12] R. Collins, R. Gross, and J. Shi, "Silhouette-based human identification from body shape and gait," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, 2002, pp. 366–371.

[13] J. Phillips, H. Moon, S. Rizvi, and P. Rause, "The FERET evaluation methodology for face recognition algorithms," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 1090–1104, Oct. 2000.

[14] C. Yam, M. Nixon, and J. Carter, "On the relationship of human walking and running: automatic person identification by gait," in *Proc. Int. Conf. Pattern Recognition*, vol. I, 2002, pp. 287–290.

[15] P. Phillips, S. Sarkar, I. Robledo, P. Grother, and K. Bowyer, "The gait identification challenge problem: Data sets and baseline algorithm," in *Proc. Int. Conf. Pattern Recognition*, vol. I, 2002, pp. 385–388.

[16] L. Lee and W. Grimson, "Gait analysis for recognition and classification," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, 2002, pp. 155–162.

[17] R. Brunelli and D. Falavigna, "Person identification using multiple cues," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, pp. 955–966, Oct. 1995.

[18] L. Hong and A. Jain, "Integrating faces and fingerprints for personal identification," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 1295–1307, Dec. 1998.

[19] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 226–239, 1998.

[20] B. Achermann and H. Bunke, "Combination of Classifiers on the Decision Level for Face Recognition," University of Bern, IAM-96-002, 1996.

[21] M. Isard and A. Blake, "A mixed-state condensation tracker with automatic modal switching," in *Proc. Int. Conf. Computer Vision*, 1998, pp. 107–112.

[22] H. Ning, L. Wang, W. Hu, and T. Tan, "Articulated model based people tracking using motion models," in *Proc. Int. Conf. Multi-Modal Interface*, 2002, pp. 383–388.

[23] L. Wang, T. Tan, W. Hu, and H. Ning, "Automatic gait recognition based on statistical shape analysis," *IEEE Trans. Image Processing*, vol. 12, pp. 1120–1131, Sept. 2003.

[24] L. Wang, H. Ning, T. Tan, and W. Hu, "Fusion of static and dynamic body biometrics for gait recognition," in *Proc. Int. Conf. Computer Vision*, vol. II, Oct. 2003, pp. 1449–1454.

[25] A. Baumberg and D. Hogg, "Generating spatiotemporal models from examples," in *Proc. Brit. Machine Vision Conf.*, 1995, pp. 413–422.

[26] D. Cunado, M. Nixon, and J. Carter, "Automatic extraction and description of human gait models for recognition purposes," *Comput/ Vis. Image Understanding*, vol. 90, no. 1, pp. 1–41, 2003.

[27] S. A. Niyogi and E. H. Adelson, "Analyzing and recognizing walking figures in XYT," in *Proc. IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition*, 1994, pp. 469–474.

[28] J. Little and J. Boyd, "Recognizing people by their gait: the shape of motion," *Videre: J. Comput. Vis. Res.*, vol. 1, no. 2, pp. 2–32, 1998.

Liang Wang received the B.Sc. degree in electrical engineering and the M.Sc. degree in video processing and multimedia communication from Anhui University, Hefei, China, in 1997 and 2000, respectively, and is currently working toward the Ph.D. degree in pattern recognition and intelligent systems at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

He has published more than 10 papers in international journals and conferences. His current research interests include computer vision, pattern recognition, digital image processing and analysis, multimedia, and visual surveillance.

Huazhong Ning received the B.Sc. degree in computer science from the University of Science and Technology of China, Hefei, in 2000 and the M.S. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

He has published several papers in national journals and international conferences. His main research interests include computer vision, human computer interaction, image processing, pattern recognition, and graphics.

Tieniu Tan (M'92–SM'97–F'03) received the B.Sc. degree in electronic engineering from Xi'an Jiaotong University, Xi'an, China, in 1984 and the M.Sc., DIC, and Ph.D. degrees in electronic engineering from Imperial College of Science, Technology and Medicine, London, U.K., in 1986, 1986, and 1989, respectively.

He joined the Computational Vision Group, Department of Computer Science, The University of Reading, Reading, U.K., in October 1989, where he was a Research Fellow, Senior Research Fellow, and Lecturer, respectively. In January 1998, he returned to China to join the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is currently a 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.

Dr. Tan 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. He serves as a 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 and is the Asia Editor of *Image and Vision Computing.*

Weiming Hu received the Ph.D. degree from Zhejiang University, Hangzhou, China.

He was a Post-Doctoral Research Fellow with the Institute of Computer Science and Technology, Founder Research and Design Center, Peking University, Peking, China, from April 1998 to March 2000. In April 2000, he joined the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, as an Associate Professor. He has published more than 40 papers in major national journals and international conference proceedings. His current research interests include visual surveillance and monitoring of dynamic scenes, recognition and filtering of Internet objectionable images, neural networks, and computer vision.