Bayesian-based Performance Prediction for Gait Recognition

Bir Bhanu and Ju Han
Center for Research in Intelligent Systems
University of California, Riverside, California 92521, USA
{bhanu,jhan}@cris.ucr.edu

Abstract

Existing gait recognition approaches do not give their theoretical or experiential performance predictions. Therefore, the discriminating power of gait as a feature for human recognition cannot be evaluated. In this paper, we first propose a kinematic-based approach to recognize human by gait. The proposed approach estimates 3D human walking parameters by performing a least squares fit of the 3D kinematic model to the 2D silhouette extracted from a monocular image sequence. Next, a Bayesian based statistical analysis is performed to evaluate the discriminating power of extracted features. Through probabilistic simulation, we not only predict the probability of correct recognition (PCR) with regard to different within-class feature variance, but also obtain the upper bound on PCR with regard to different human silhouette resolution. In addition, the maximum number of people in a database is obtained given the allowable error rate. This is extremely important for gait recognition in large databases.

1 Introduction

Model-based object recognition is concerned with searching for a match: how to associate components of the given data with corresponding parameters of the object model [2]. From this viewpoint, approaches can be classified as global matching or feature matching. Global matching approaches consider finding a transformation from a model to an image while feature matching approaches involve establishing a correspondence between local features extracted from the given data and corresponding local features of the object model.

Boshra and Bhanu [1] present a method for predicting fundamental performance of object recognition. They assume that both scene data and model objects are represented by 2D point features and a data/model match is evaluated using a vote-based criterion. Their method considers data distortion factors such as uncertainty, occlusion, and clutter, in addition to model similarity. This is unlike previous approaches, which consider only a subset of these factors. However, their assumptions make their method only applicable to feature matching and not to global matching.

In our proposed approach of human recognition by kinematic-based gait analysis, we use global matching because we only have the global human silhouette information before matching. The detailed information for different body parts is obtained after matching. Next, we carry out Bayesian based statistical analysis to evaluate the discriminating power of various features. We address the prediction problem in the context of an object recognition task as follows: (1) scene data are represented by 2D regions where the region pixels are discretized at some resolution, and model objects are represented by 3D volumes; (2) an instance of a model object in the scene data is assumed to be obtained by applying some 3D to 2D transformation to the object; (3) the matching criterion is Bayesian theory.

2 Motivation and Contributions

Current human recognition methods, such as fingerprints, face or iris biometrics, generally require a cooperative subject, views from certain aspects and physical contact or close proximity. These methods can not reliably recognize non-cooperating individuals at a distance in real-world changing environmental conditions. Moreover, in many applications of personnel identification, many established biometrics can be obscured. Gait, which concerns recognizing individuals by the way they walk, can be used as a biometric without the above-mentioned disadvantages.

In recent years, some approaches have already been employed in automatic gait recognition (i.e., human recognition by gait). Niyogi and Adelson [9] make an initial attempt in a spatiotemporal (XYT) volume. They first find the bounding contours of the walker, and then fit a simplified stick model on them. A characteristic gait pattern in XYT is generated from the model parameters for recognition. Little and Boyd [7] propose a model-free approach making no attempt to recover a structural model of human motion. Instead they describe the shape of the motion with a set of features derived from moments of a dense flow distribution. Similarly, He and Debrunner's [3] approach detects a sequence of feature vectors based on Hu's moments of motion segmentation in each frame, and the individual is recognized from the feature vector sequence using hidden Markov models. To avoid a feature extraction process which may reduce reliability, Murase and Sakai [8] propose a template matching

Proceedings of the Workshop on Motion and Video Computing (MOTION'02)
0-7695-1860-5/02 $17.00 © 2002 IEEE

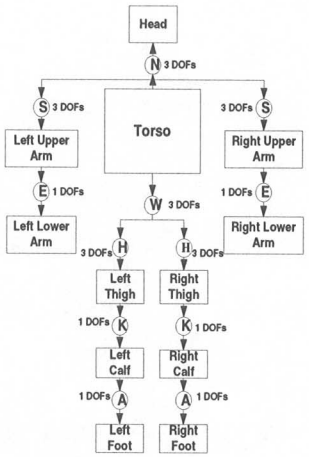

Figure 1. 3D Human Kinematic Model.

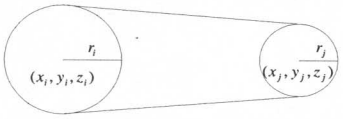

Figure 2. Body part geometric representation.

method to calculate the spatio-temporal correlation in a parametric eigenspace representation for gait recognition. Huang et al. [5, 4] extend this approach by combining canonical space transformation (CST) with eigenspace transformation (EST) for feature selection.

However, existing gait recognition approaches only consider human walking frontoparallel to the image plane. Moreover, none of the existing gait recognition approaches give their theoretical or experiential performance prediction. Therefore, we cannot evaluate the discriminating power of gait as a feature for human recognition. In this paper, we propose a kinematic-based approach to recognize human by gait, and carry out Bayesian based statistical analysis to predict recognition performance. The proposed approach estimates 3D human walking parameters by performing a least squares fit of the 3D kinematic model to the 2D silhouette extracted from a monocular image sequence. The gait features are then generated from the estimated model parameters for human recognition. Our approach eliminates the assumption of human walking frontoparallel to the image plane, which is desirable in many gait recognition applications.

3 Technical Approach

3.1 Human Kinematic Model

A human body is considered as an articulated object, consisting of a number of body parts. The body model adopted here is shown in Figure 1, where a circle represents a joint and a rectangle represent a body part (N: neck, S: shoulder, E: elbow, W: waist, H: hip, K: knee, and A: ankle). Most joints and body part ends can be represented as spheres, and most body parts can be represented as cones. The whole human kinematic model is represented as a set of cones connected by spheres [6]. Figure 2 shows that most of the body parts can be approximated well in this manner. However, the head is approximated only crudely by a sphere and the torso is approximated by a cylinder with two spheroid ends.

3.2 Matching 3D Model with 2D Silhouette

The matching procedure determines a parameter vector $\mathbf{x}$ so that the proposed 3D model fits the given 2D silhouette as well as possible. For that purpose, two chained transformations transform human body local coordinates (x, y, z) into image coordinates (x', y')[12]. The first transformation transforms local coordinates into camera coordinates; while the second transformation projects camera coordinates into image coordinates.

Each 3D human body part is modeled by a cone with two spheres $\mathbf{s}_i$ and $\mathbf{s}_j$ at its ends, as shown in Figure 2 [6]. Each sphere $\mathbf{s}_i$ is fully defined by 4 scalar values, (x_i, y_i, z_i, r_i), which define its location and size. Given these values for two spheroid ends (x_i, y_i, z_i, r_i) and (x_j, y_j, z_j, r_j) of a 3D human body part model, its projection $P_{(ij)}$ onto the image plane is the convex hull of the two circles defined by (x'_i, y'_i, r'_i) and (x'_j, y'_j, r'_j).

If the 2D human silhouette is known, we may find the relative 3D body parts locations and orientations with the knowledge of camera parameters. We propose a method to perform a least squares fit of the 3D human model to the 2D human silhouette. That is, to estimate the set of sphere parameters $\mathbf{x} = \{\mathbf{x}_i : (x_i, y_i, z_i, r_i)\}$ by choosing $\mathbf{x}$ to minimize

$$error(\mathbf{x}; I) = \sum_{x', y' \in I} (P_{\mathbf{x}}(x', y') - I(x', y'))^2, \qquad (1)$$

where I is the silhouette binary image, $P_{\mathbf{x}}$ is the binary projection of the 3D human model to image plane, and x', y' are image plane coordinates.

3.3 Model Parameter Selection

Human motion is very complex due to so many degrees of freedom (DOFs). To simplify the matching procedure, we use the following reasonable assumptions: (1) the camera is stationary; (2) people are walking before the camera at a distance; (3) people are moving in a constant direction; (4) the swing direction of arms and legs parallels to the moving direction. According to these assumptions, we do not need to consider the waist joint, and only need to consider one DOF for each other joint. Therefore, the elements of the parameter vector of the 3D human kinematic model are defined as: (a) Radius $r_i(11)$: torso(3), shoulder, elbow, hand, hip, knee, ankle, toe, and head; Length $l_i(9)$: torso, inter-shoulder, inter-hip, upper arm, forearm, thigh, calf, foot, and neck; (b) Location $(x, y)(2)$; Angle $\theta_i(11)$: neck, left upper arm, left forearm, right

Proceedings of the Workshop on Motion and Video Computing (MOTION'02)
0-7695-1860-5/02 $17.00 © 2002 IEEE

upper arm, right forearm, left thigh, left calf, left foot, right thigh, right calf, and right foot. With 33 stationary and kinematic parameters, the projection of the human model can be completely determined.

3.4 Silhouette Extraction

Assuming that people are the only moving objects in the scene, they can be extracted by a simple background subtraction method. Notice that an area cast into shadow often results in a significant change in intensity without much change in chromaticity. Given a video sequence containing moving people and the corresponding background image, for each frame I_i in the sequence, the color value difference $\Delta \mathbf{p}_i(x,y) = ||\mathbf{p}_i(x,y) - \mathbf{p}_b(x,y)||$ is computed for each pixel, where $\mathbf{p}_i(x,y)$ and $\mathbf{p}_b(x,y)$ are RGB color values of the pixel at (x,y) in the ith frame and background image, respectively. The chromaticity is computed as

$$\begin{aligned} r_c(x,y) &= r(x,y)/(r(x,y)+g(x,y)+b(x,y)) \\ g_c(x,y) &= g(x,y)/(r(x,y)+g(x,y)+b(x,y)). \end{aligned}$$

$$\begin{aligned} \text{We have} \quad \Delta r_{ci}(x,y) &= |r_{ci}(x,y) - r_{cb}(x,y)| \\ \text{and} \quad \Delta g_{ci}(x,y) &= |g_{ci}(x,y) - g_{cb}(x,y)|. \end{aligned}$$

Given thresholds t_1 and t_2, if

$$(\Delta \mathbf{p}_i(x,y) > t_1) \wedge ((\Delta r_{ci}(x,y) > t_2) \vee (\Delta g_{ci}(x,y) > t_2))$$

the pixel at (x,y) is determined to be part of the moving objects; otherwise, it is part of the background.

After the silhouette has been cleaned by a preprocessing procedure, its height, width and centroid can be easily extracted for motion analysis. In addition, the moving direction of the walking person is determined as follows

$$\theta = \begin{cases} tan^{-1}\frac{f(h_1-h_N)}{h_1 y_N - h_N y_1}, & \text{if } y_1 > y_N; \\ tan^{-1}\frac{f(h_1-h_N)}{h_1 y_N - h_N y_1} + \pi, & \text{otherwise.} \end{cases} \quad (2)$$

where f is the camera focal length, y_1 and y_N are the horizontal centroid of the silhouette in the first and Nth frame, and h_1 and h_N are the height of the silhouette in the first and Nth frame.

3.5 Stationary Parameter Estimation

The stationary parameters include body part length and joint radius. Notice that human walking is a cyclic motion, so a video sequence can be divided into motion cycles and studied separately. In each walking cycle, the silhouette with minimum width means that the person stands straight and that means the most occlusion; the silhouette with maximum width means the least occlusion and, therefore, it is more reliable.

To estimate the stationary parameters, we first select some key frames (4 frames in our experiments) which contain more reliable silhouettes, and then perform matching procedure on the key frames as a whole. The corresponding feature vector thus includes 20 common stationary parameters and 13*4 individual kinematic parameters. Next, we first initialize these parameters according to the human statistical information. Then, the set of parameters is estimated from these initial parameters by choosing a parameter vector $\mathbf{x}$ to minimize the least square error in equation (1) with respect to the same kinematic constraints.

After the matching algorithm is converged, the estimated stationary parameters so obtained are used for kinematic parameter estimation of other frames. At the same time, the estimated kinematic parameters of key frames are used for prediction. Because even the same person might walk at different speed, we normalize the estimated kinematic parameters of each walking cycle to a fix-length walking cycle, and the gait features are generated from the normalized walking cycle.

4 Recognition Performance Prediction

In this paper, we only use features from stationary parameters for gait recognition. In the abovementioned stationary parameters, radius parameters will be different if the same person is in different clothes, and are thus not reliable for recognition. Similarly, inter-shoulder and inter-hip length parameters are not reliable because people usually walk within some angle along the direction pedicular to the camera axis. The head region depends on the hair style, which will change if the view changes, and the head representation in our model (sphere) is not precise in some cases, so the estimated neck length is also not reliable. Therefore, the feature vector selected for human recognition in our approach includes 6 elements: torso length, upper arm length, forearm length, thigh length, calf length, and foot length, which are not sensitive for recognizing human with different clothes. In this paper, we consider uncertainties for feature vectors in two ways: the ideal case - uncertainties only from different resolution, and the general case - uncertainties from various factors.

4.1 Body Part Length Distribution

To predict the performance of recognizing human from body part lengths, we have to know the prior length distributions of body parts over human population. The data are called static anthropometric data shown in Table 1 [10]. Although the data are surveyed in the British population, the predicted performance on it is still applicable in some sense. In general, the mean of body part lengths will change but the standard deviation will not change a lot in different populations. Assuming that men and women have the same population, the overall distributions for each of the body part lengths are obtained. In this paper, we only consider that the body part lengths are independently distributed due to the absence of statistical knowledge of their correlation.

Proceedings of the Workshop on Motion and Video Computing (MOTION'02)
0-7695-1860-5/02 $17.00 © 2002 IEEE

body part length	μ men	σ men	μ women	σ women
torso length	595	32	555	31
upper arm length	365	20	360	17
forearm length	475	21	430	19
thigh length	495	32	480	30
calf length	545	32	500	27
foot length	265	14	235	12

Table 1. Anthropometric estimates for British adults 19-65 years (all lengths in millimeters).

4.2 Performance Prediction in the General Case

In the general case, uncertainties of features come from various sources: image quantization error, camera calibration error, silhouette segmentation error, matching error, and body part occlusion. To completely model the uncertainties of 3D body part lengths, we have to model all the above-mentioned factors. This is a challenging task because it is difficult to mathematically find the distribution functions of uncertainties for all these factors. A reasonable approach is to estimate the uncertainties on ground truth data, i.e., training data. Assuming that feature vectors obtained from a feature extraction approach for a person are normally distributed in the given feature space, we can easily obtain the within-class variance from the experimental results on the training data. Then, the obtained within-class variance can be used to predict the recognition performance of this approach.

According to Bayes decision theory, an unknown feature vector $\mathbf{x}$ is assigned to class ω_i if $P(\omega_i|\mathrm{x}) > P(\omega_j|\mathrm{x})\ \forall j \neq i$ [11]. Let $g_i(\mathrm{x}) = \ln(p(\mathrm{x}|\omega_i)P(\omega_i))$, this decision test becomes classifying $\mathbf{x}$ in ω_i if $g_i(\mathrm{x}) > g_j(\mathrm{x})\ \forall j \neq i$.

Assuming the feature vector $\mathbf{x}$ for a person ω_i is normally distributed in l-dimensional feature space, the likelihood functions of ω_i with respect to $\mathbf{x}$ follow

$$p(\mathbf{x}|\omega_i) = \frac{1}{(2\pi)^{\frac{1}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i)\right)$$

for $i = 1, ..., M$, where $\mu_i = E[\mathbf{x}]$ is the mean value of the ω_i class and Σ_i is the $l \times l$ covariance matrix defined as $\Sigma_i = E[(\mathbf{x}-\mu_i)(\mathbf{x}-\mu_i)^T]$. For simplicity, we assume that $\Sigma_i = \sigma I$ for all i, i.e., each of the independent features has an identical Gaussian distribution. Therefore, maximum $g_i(\mathrm{x})$ implies minimum Euclidean distance: $d_E = ||\mathbf{x}-\mu_i||$. Thus, feature vectors are assigned to classes according to their Euclidean distances from the respective mean vectors.

With the body part length distribution in Table 1 and the within-class standard deviation σ of the features obtained from a feature extraction approach, we can predict its *probability of correct recognition* (PCR) with regard to the number of classes (people) in the database.

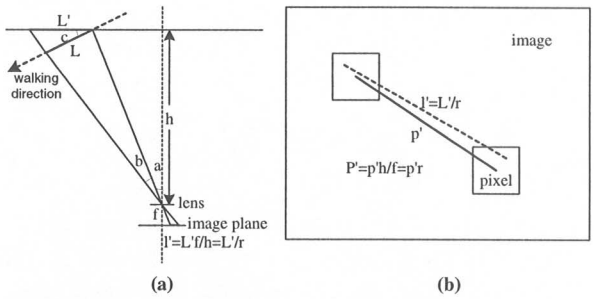

Figure 3. Uncertainty computation for the given silhouette resolution r (in millimeters/pixel), the body part length L, view angles a and b, and the walking direction c.

4.3 Upper Bound on PCR

We have considered the uncertainties in the general case, which are dependent on feature extraction approaches. Although the predicted performance indicates the discriminant power of features extracted by different approaches, and these approaches can therefore be compared, we still do not know the upper bound on PCR which can be achieved. In the ideal case, image quantization errors, i.e., the human silhouette resolution, is the only source of uncertainties. By analyzing the uncertainties given a fixed silhouette resolution, we can obtain the upper bound on PCR with regard to the number of classes (people) in the database.

Given the silhouette resolution r, we can compute the corresponding uncertainty from the body part length L, view angles a and b, and the walking direction c through two steps as shown in Figure 3. The first step is projecting the 3D length L to length l' in the 2D continuous plane. We obtain the projection of L on the plane at depth h which is perpendicular to the camera axis as follows:

$$L' = L(\cos c + \sin c \tan(a+b)). \tag{3}$$

Figure 3(a) only shows the case where $a > 0$, $b > 0$ and $c > 0$. We can easily derive the same equation in other cases. Then the corresponding length of L in the continuous image plane can be computed from the following equation:

$$l' = L'f/h = L'r, \tag{4}$$

where f is the camera focus length.

The second step is the image quantization step as shown in Figure 3(b). For every 2D point falling into a box in the continuous image plane, it's location is represented by the center location of the box in the discrete image plane. Therefore, the corresponding length of L in the discrete image plane is the discrete value p'. From Equations 3 and 4, we can obtain the following results

$$P' = p'h/f = p'r, \tag{5}$$

Proceedings of the Workshop on Motion and Video Computing (MOTION'02)
0-7695-1860-5/02 $17.00 © 2002 IEEE

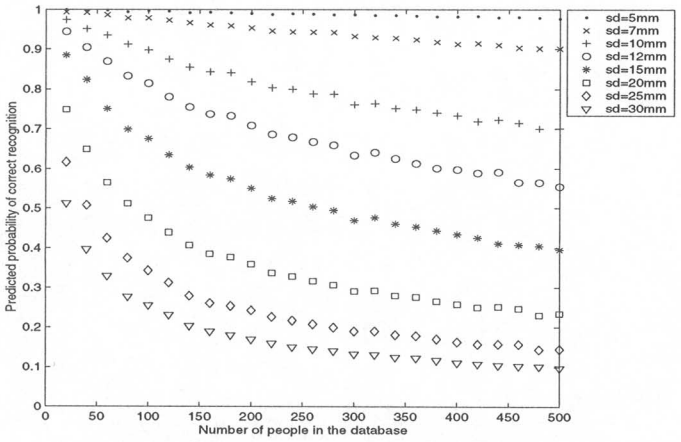

Figure 4. Predicted PCR with regard to different database size and different within-class standard deviation values of the extracted features (general case).

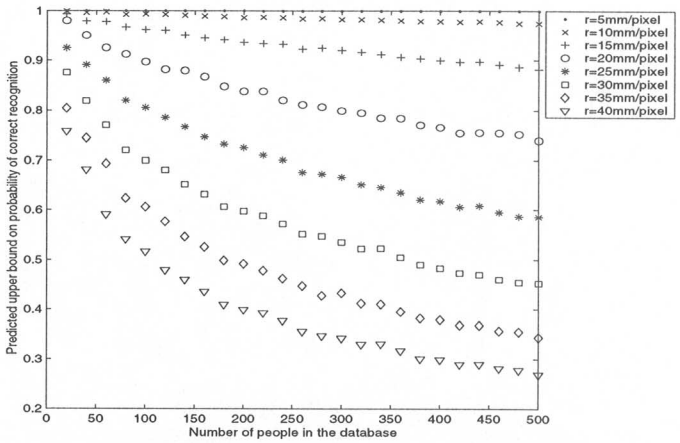

Figure 5. Predicted upper bound on PCR with regard to different database size and different human silhouette resolution values (ideal case).

$$P = \frac{P'}{\cos c + \sin c \tan(a+b)} = \frac{p'r}{\cos c + \sin c \tan(a+b)}, \quad (6)$$

where P is the corresponding length of p' in 3D space and P' is the projection of P on the plane at depth h which is perpendicular to the camera axis. Therefore, the overall error in Figure 3 is $P - L$. Considering $h \gg L$ in our applications, we have $b \approx 0$, and Equation (6) becomes

$$P = p'r/(\cos c + \sin c \tan a). \quad (7)$$

Assuming the elements in the feature vector are identically independently distributed, the Euclidean distance classification criteria is still effective.

Assuming that the quantization error is uniformly distributed in the $r \times r$ area, view angle a is uniformly distributed from $-45°$ to $45°$ of arc, and walking direction c is uniformly distributed from $-30°$ to $30°$ of arc, we can predict the recognition performance with regard to the number of classes (people) in the database through a simulation approach. The prediction results are upper bounds on PCR with regards to different human silhouette resolution values.

5 Experimental Results

5.1 Performance Prediction Results

In our experiments, the performance prediction results are obtained through simulation approaches. First we randomly generate the body part lengths of M classes (people) according to the distribution of different body part lengths. Next, for each of the M classes, we randomly generate N instances for this class according to the uncertainties in Section 4.2 or Section 4.3. Finally, we obtain the recognition rate in the current experiment. After this experiment has been repeated for K times, we can obtain the average recognition rate. If $M * K$ are large enough ($M = 100$ and $K = 100$ in our experiments), this average recognition rate can be viewed as the predicted PCR in the general case (Figure 4), and upper bounds on PCR in the ideal case (Figure 5). From these prediction results, we can find the corresponding maximum number of people in a database given the allowable error rate. Table 2 shows the corresponding resolution (mm/pixel) for a 1675 mm (population average height) person occupying different vertical portions of the frame with different video formats. It is shown that most of these resolutions are good enough for human recognition in databases of less than 500 people.

Human Silhouette Occupancy	VHS 240 lines	Digital Video 480 lines	High Definition 1080 lines
100% of frame	6.98	3.49	1.55
75% of frame	9.31	4.65	2.07
50% of frame	13.96	6.98	3.10
25% of frame	27.92	13.96	6.20

Table 2. Resolution (mm/pixel) for a 1675 mm (population average height) person occupying different vertical portions of the frame with different video formats.

Our prediction results are based on the assumption that the selected length features are independently distributed with an identical Gaussian distribution. This assumption may not accurately reflect different types of perturbations. In the future, We will investigate the real feature distribution under different types of perturbations.

5.2 Recognition Results on Real Data

The video data used in our experiments are real human walking data recorded in an outdoor environment, and there is only one walking person at the same time.

Proceedings of the Workshop on Motion and Video Computing (MOTION'02)
0-7695-1860-5/02 $17.00 © 2002 IEEE

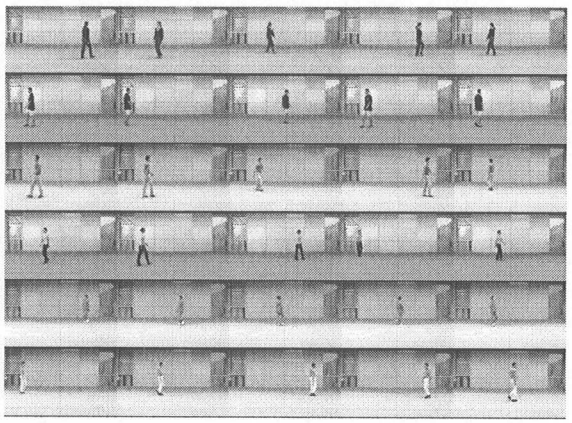

Figure 6. Sample sequences in our database.

Eight different people walk along different directions (within $[-45°, 45°]$ along the image plane). The size of image frames is 180×240. In our experiments, we first manually divide video data into single-cycle sequences, and then select 15 sequences from each person: 10 sequences for training and 5 sequences for testing. Figures 6 shows sequences in our gait database.

The least square matching algorithm is implemented using a Genetic algorithm. The fitness function is computed from the matching error in Equation (1). In our experiments, our approach achieves 60% recognition rate on the training dataset using the Leave-One-Out method. The performance on the testing data is 42% recognition rate. We also compute the average standard deviation for each person in the database which is 20 mm, and the corresponding predicted PCR is 87%. The correct recognition rate in our approach is much lower than this PCR because the PCR is computed on the data distributed according to Table 1 while the data in our database are not well distributed due to the small data size, i.e., they have more similarity. The human silhouette resolution in our database varies from 20 to 30 mm/pixel, and the corresponding predicted upper bound on PCR in the ideal case is from 94.67% to 98.80%. The predicted PCR (87%) is lower than the upper bound because our feature extraction approach introduces several additional uncertainties such as camera calibration error, silhouette segmentation error, matching error, and body part occlusion.

Note that the use of binary silhouette to fit 3D model suffers from ambiguity as a result of body parts self-occlusion, and the use of least squares makes it sensitive to noise in the silhouette. However, this problem can be solved by considering the correlation between adjacent frames.

6 Conclusions

In this paper, we proposed a Bayesian based statistical analysis to evaluate the discriminating power of extracted features. Through probabilistic simulation, we not only obtain the PCR for our approach, but also obtain the upper bound on PCR with regard to different human silhouette resolution in ideal cases. The obtained PCR for our approach is lower than the upper bound because our feature extraction approach introduces several additional uncertainties such as image quantization error, camera calibration error, silhouette segmentation error, matching error, and body part occlusion. Through the theoretical analysis, we obtain the upper bound on PCR for the given silhouette resolution, and can accordingly improve the recognition performance by reducing error introduced. In addition, we obtain the maximum number of people in the database given the allowable error rate. This will guide future research for gait recognition in large databases.

Acknowledgment

This work was supported in part by grants F49620-97-1-0184, F49620-02-1-0315 and DAAD19-01-0357; the contents and information do not necessarily reflect the position or policy of U.S. Government.

References

[1] M. Boshra and B. Bhanu. Predicting performance of object recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(9):956–969, 2000.

[2] W. Grimson. *Object recognition by computer: the role of geometric constraints.* The MIT Press, 1990.

[3] Q. He and C. Debrunner. Individual recognition from periodic activity using hidden markov models. in *Proc. IEEE Workshop on Human Motion*, pages 47–52, 2000.

[4] P. Huang. Automatic gait recognition via statistical approaches for extended template features. *IEEE Trans. Systems, Man and Cybernetics, Part B*, 31(5):818–824, 2001.

[5] P. Huang, C. Harris, and M. Nixon. Recognizing humans by gait via parameteric canonical space. *Artificial Intelligence in Engineering*, 13:359–366, 1999.

[6] M. Lin. Tracking articulated objects in real-time range image sequences. in *Proc. International Conference on Computer Vision*, pages 648–653, 1999.

[7] J. Little and J. Boyd. Recognizing people by their gait: the shape of motion. *Videre: Journal of Computer Vision Research*, 1(2):1–32, 1998.

[8] H. Murase and R. Sakai. Moving object recognition in eigenspace representation: gait analysis and lip reading. *Pattern Recognition Letters*, 17(2):155–62, 1996.

[9] S. Niyogi and E. Adelson. Analyzing and recognizing walking figures in xyt. in *Proc. IEEE Conference on CVPR*, pages 469–474, 1994.

[10] S. Pheasant. *Bodyspace: anthropometry, ergonomics and design.* Taylor & Francis, 1986.

[11] S. Theodoridis and K. Koutroumbas. *Pattern Recognition.* Academic Press, 1998.

[12] S. Wachter and H.-H. Nagel. Tracking of persons in monocular image sequences. in *Proc. IEEE Workshop on Nonrigid and Articulated Motion*, pages 2–9, 1997.

Proceedings of the Workshop on Motion and Video Computing (MOTION'02)
0-7695-1860-5/02 $17.00 © 2002 **IEEE**