A HIDDEN MARKOV MODEL BASED FRAMEWORK FOR RECOGNITION OF HUMANS FROM GAIT SEQUENCES

Aravind Sundaresan, Amit RoyChowdhury, Rama Chellappa

Centre for Automation Research, and Department of Electrical and Computer Engineering
University of Maryland, College Park, MD 20742

ABSTRACT

In this paper we propose a generic framework based on Hidden Markov Models (HMMs) for recognition of individuals from their gait. The HMM framework is suitable, because the gait of an individual can be visualized as his adopting postures from a set, in a sequence which has an underlying structured probabilistic nature. The postures that the individual adopts can be regarded as the states of the HMM and are typical to that individual and provide a means of discrimination. The framework assumes that, during gait, the individual transitions between N discrete postures or states but it is not dependent on the particular feature vector used to represent the gait information contained in the postures. The framework, thus, provides flexibility in the selection of the feature vector. The statistical nature of the HMM lends robustness to the model. In this paper we use the binarized background-subtracted image as the feature vector and use different distance metrics, such as those based on the L_1 and L_2 norms of the vector difference, and the normalized inner product of the vectors, to measure the similarity between feature vectors. The results we obtain are better than the baseline recognition rates reported before.

1. INTRODUCTION

Biometrics can be a powerful cue for reliable automated person identification and there exist several established biometric-based identification techniques including fingerprint and hand geometry methods, speaker identification, face recognition and iris identification. However, the applicability of all these methodologies is usually restricted to controlled environments or require cooperation of the subject. We, therefore, need to explore biometric signatures which can be obtained *non-invasively* from a distance. Gait is one such biometric which is currently being explored for purposes such as identification. We know from experience that people often recognize others by simply observing the way they walk implying that body shape and dynamics are sufficiently distinct across humans.

It has been observed that gait can be modelled as a transition across states, where each state is exemplified by a typical feature vector or *exemplar*. In this paper we propose a general framework that employs exemplar-based HMMs to characterize an individual's gait and thereafter recognize the individual from his gait. HMMs have been used in speech modelling and [1] provides a good tutorial on HMMs. The use of *temporal templates*, which capture motion of each pixel in the frame, has been proposed for recognition of both human actions [2] and humans from their gait [3]. HMMs have also been used to recognize human actions [4, 5, 6]. Exemplars have been used in learning probabilistic models in [7] and tracking in [8]. Our objective is to recognize an individual from a video sequence of the individual walking in a fronto-parallel pose. We represent the structural aspect of the person by using typical feature vectors corresponding to different "states", and the dynamical aspect of the individual's gait by modelling the transition between these "states" using the transition matrix like in [9]. We integrate these two components to train models for representation and identification. Additionally, we propose a probability distribution for the observations based on the exemplars. This framework affords us flexibility in our choice of feature vectors, and suitable distance metrics corresponding to the feature vectors that we choose. We provide algorithms to train the exemplars as a function of all the feature vectors unlike [9] which selects just N frames as the exemplars and trains the *Feature-to-Exemplar distance* vectors instead. This step makes our algorithm much less susceptible to noise in the feature vectors. In our experiments, we have used video sequences from the USF database[10]. We compare the performance of our algorithm to the baseline algorithm proposed in [11].

2. OVERVIEW OF THE HMM FRAMEWORK

Let the database consists of video sequences of P persons. The HMM for the p^{th} person is given by $\lambda_p = (A_p, B_p, \pi_p)$ with N number of states. The model, λ_p, is trained using the observation sequence for the p^{th} person, i.e. the sequence of feature vectors $\mathcal{O}_p = \{\mathbf{O}_1^p, \mathbf{O}_2^p, \ldots, \mathbf{O}_{T_p}^p\}$, where T_p is the number of frames in the video sequence of the p^{th} person. A_p is the transition matrix, and π_p is the initial distribution. The B_p parameter comprises of the probability distributions for a feature vector conditional on the state index, i.e., the set $\{P_1^p(.), P_2^p(.), \ldots, P_N^p(.)\}$ (see (1)). The probability distributions are defined in terms of *exemplars*, where the j^{th} exemplar is a typical realization of the j^{th} state. The exemplars for the p^{th} person are given by $\mathcal{E}_p = \{\mathbf{E}_1^p, \mathbf{E}_2^p, \ldots, \mathbf{E}_N^p\}$. Henceforth, the superscript denoting the index of the person is dropped for simplicity. The motivation behind using an exemplar-based model is that the recognition can be based on the distance measure between the observed feature vector and the exemplars. The distance metric is evidently a key factor in the performance of the algorithm. $P_j(\mathbf{O}_t)$ is defined as a function of $D(\mathbf{O}_t, \mathbf{E}_j)$, the distance of the feature vector $\mathbf{O}_t$ from the j^{th} exemplar.

$$P_j(\mathbf{O}_t) = \alpha e^{-\alpha D(\mathbf{O}_t, \mathbf{E}_j)} \tag{1}$$

During the *training* phase, a model is built for all the subjects in the gallery, indexed by $p = 1, 2, \ldots, P$. An initial estimate of

Supported by the DARPA/ONR grant N00014-03-1-0520

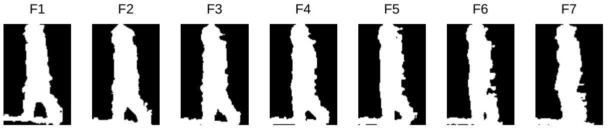


Fig. 1. Part of an Observation Sequence

$\mathcal{E}_p$ and λ_p is formed from $\mathcal{O}_p$, and these estimates are refined iteratively using Expectation-Maximisation. Note that B is completely defined by $\mathcal{E}$ if α is fixed beforehand. We can iteratively refine estimates of A and π by using the Baum-Welch algorithm and keeping $\mathcal{E}$ fixed. The algorithm to refine estimates of $\mathcal{E}$ is determined by the choice of the distance metric. Given a Gallery $\mathcal{L} = \{\lambda_1, \lambda_2, \dots, \lambda_P\}$, we identify a probe sequence of length T, $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_T\}$ traversing an unknown path $\mathcal{Q} = \{q_1, q_2, \dots, q_T\}$, q_t being the state index at time t, as

$$ID = \arg_p \max_{\mathcal{Q},p} \Pr[\mathcal{Q}|\mathcal{X}, \lambda_p]. \qquad (2)$$

3. METHODOLOGY

The feature vector we use in the experiments is the binarized version of the background subtracted images provided in the USF database. The images are scaled and aligned to the centre of the frame as in Fig. 1 which features part of a sequence of feature vectors. We describe in this section the methods used to obtain initial estimates the HMM parameters, the training algorithm and finally and identification from a probe sequence.

3.1. Initial Estimate of HMM Parameters

In order to obtain a good estimate of the exemplars and the transition matrix, we first obtain an initial estimate of an ordered set of exemplars from the sequence and the transition matrix and iteratively refine the estimate. The initial estimate for the exemplars, $\mathcal{E}^0 = \{\mathbf{E}_1^0, \mathbf{E}_2^0, \dots, \mathbf{E}_N^0\}$ is such that the only transitions allowed are from the j^{th} state to either the j^{th} or the $(j \bmod N + 1)^{th}$ state. A corresponding initial estimate of the transition matrix, A^0 (with $A^0_{j,j} = A^0_{j,j \bmod N+1} = 0.5$, and all other $A^0_{j,k} = 0$) is also obtained. The initial probabilites π_j are set to be equal to $1/N$.

We observe that the gait sequence is quasi-periodic and we use this fact to obtain the initial estimate $\mathcal{E}^0$. We can divide the sequence into “cycles”, where a cycle is defined as that segment of the sequence bounded by silhouettes where the subject has arms by his side and legs approximately aligned with each other. We can further divide each cycle into N temporally adjacent clusters of approximately equal size. We visualize the frames of the j^{th} cluster of all cycles to be generated from the j^{th} state. Thus we can get a good initial estimate of $\mathbf{E}_j$ from the feature vectors belonging to the j^{th} cluster. For example, assume that the training sequence is given by $\mathcal{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_T\}$. We can partition the sequence into K cycles, with the k^{th} cycle given by frames in the set $\mathcal{Y}_k = \{\mathbf{Y}_{S_k}, \mathbf{Y}_{S_k+1}, \dots, \mathbf{Y}_{S_k+L_k-1}\}$, where S_k and L_k are the index of the first frame of the k^{th} cycle, and the length of the k^{th} cycle respectively. We define the first cluster to comprise of frames with indices $S_k, S_k+1, \dots, S_k + \frac{1}{2}L_k/N, S_k + L_k - \frac{1}{2}L_k/N, S_k + L_k - \frac{1}{2}L_k/N + 1, \dots, S_k + L_k - 1$. The j^{th} cluster ($j = 2, 3, \dots, N$) is defined to comprise of frames with indices $S_k + (j - \frac{3}{2})L_k/N, S_k + (j - \frac{3}{2})L_k/N + 1, \dots, S_k + (j - \frac{1}{2})L_k/N$.

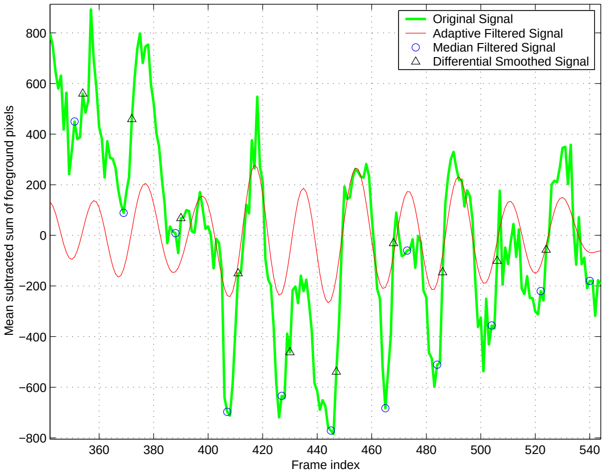


Fig. 2. Cycle boundaries estimated using an adaptive filter

We need to robustly estimate the cycle boundaries so that we can partition the sequence into N clusters and obtain the initial estimates of the exemplars. If the sums of the foreground pixels of each image are plotted with respect to time, then, as per our definition of a cycle, the minimas should correspond to the cycle boundaries. We denote the sum of the foreground pixels of the silhouette in the n^{th} frame as $s[n]$. This signal is noisy and may contain several spurious minima. However we can exploit the quasi-periodicity of the signal and filter the signal to remove the noise before identifying the minima. Methods such as median filtering or differential smoothing of $s[n]$ are not very robust as they do not take into account the frequency of the gait. A more robust method would be to analyse $s[n]$ in the frequency or the time-frequency domain, using band-pass or low-pass filters or the Short Time Fourier Transform (STFT).

The specifications of the band-pass filter are such as to allow frequencies that are typical for very-slow-to-fast walk. The video is captured at 30 frames per second, and the sampling frequency, $f_s = 1/30$ and $T_s = 30$. The maximum gait frequency is assumed to be $f_m = 0.1$ corresponding to a cycle period of $T_m = 10$. A Hamming window of length L is used. The extended sequence $x[n]$ is obtained by symmetrically extending $s[n]$ in both directions by $L/2$. Therefore the sequence $x[n]$ has length $M = N + L$. The resultant sequence is filtered using a band-pass filter (with upper cut-off frequency $f_{uc} = f_m$), in both directions to remove phase delay. The distances between the minimas of the filtered sequence lead us to an estimate of the cycle period. The cycle frequency is estimated as the inverse of the median of cycle periods. Using this revised estimate of the frequency of the gait, $\hat{f}$, a new filter is constructed with upper cut off frequency $f_{uc} = \hat{f} + 0.02$. For long sequences, STFT techniques could be used instead of straightforward filtering in order to account for a slowly varying frequency of gait. Fig. 2 illustrates the performance of several algorithms in identifying the cycle boundaries. A manual examination of all the sequences in the Gallery revealed a 100% detection rate with hardly any false detection of cycle bouldaries. We also observed from the the initial exemplars obtained using the cycle registration results from the filtering method that the frequency based filtering method is more accurate and robust compared to the other methods.

3.2. Training the HMM Parameters

The iterative refining of the estimates is performed in two steps. In the first step, a Viterbi evaluation [1] of the sequence is performed using the current values for the exemplars and the transition matrix. Thus feature vectors are clustered according to the most likely state they originated from. The exemplars for the states are newly estimated from these clusters. Using the current values of the exemplars, $\mathcal{E}^{(i)}$ and the transition matrix, $A^{(i)}$, Viterbi decoding is performed on the sequence $\mathcal{Y}$ to obtain the most probable path $\mathcal{Q} = \{q_1^{(i)}, q_2^{(i)}, \ldots, q_T^{(i)}\}$, where $q_t^{(i)}$ is the state at time t. Thus the set of observation indices, whose corresponding observation is estimated to have been generated from state j is given by $\mathcal{T}_j^{(i)} = \{t : q_t^{(i)} = j\}$. We now have a set of frames for each state and we would like to select the exemplars so as to maximise the probability in (3). If we use the definition in (1), (4) follows.

$$\mathbf{E}_j^{(i+1)} = \arg_{\mathbf{E}} \max \prod_{t \in \mathcal{T}_j^{(i)}} P(\mathbf{Y}_t|\mathbf{E}) \qquad (3)$$

$$\mathbf{E}_j^{(i+1)} = \arg_{\mathbf{E}} \min \sum_{t \in \mathcal{T}_j^{(i)}} D(\mathbf{Y}_t, \mathbf{E}) \qquad (4)$$

The actual method for minimising the distance in (4) however depends on the distance metric used. We have experimented with three different distance measures, namely the Euclidean (EUCLID) distance, the inner product (IP) distance, and the sum of absolute difference (SAD) distance which are given by (5), (6), and (7) respectively. Note that though $\mathbf{Y}_t$ and $\mathbf{E}$ are 2-dimensional images, they are represented as vectors of dimension $D \times 1$ for ease of notation. $\mathbf{1}_{D\times 1}$ is a vector of D ones.

$$D_{EUCLID}(\mathbf{Y}, \mathbf{E}) = (\mathbf{Y} - \mathbf{E})^T (\mathbf{Y} - \mathbf{E}) \qquad (5)$$

$$D_{IP}(\mathbf{Y}, \mathbf{E}) = 1 - \frac{\mathbf{Y}^T \mathbf{E}}{\sqrt{\mathbf{Y}^T \mathbf{Y} \mathbf{E}^T \mathbf{E}}} \qquad (6)$$

$$D_{SAD}(\mathbf{Y}, \mathbf{E}) = |\mathbf{Y} - \mathbf{E}|^T \mathbf{1}_{D\times 1} \qquad (7)$$

The equations for updating the j^{th} element of the exemplars in the EUCLID distance, IP distance and the SAD distance cases are presented in (8), (9) and (10) respectively. $\tilde{\mathbf{Y}}$ denotes the normalized vector $\mathbf{Y}$ and $|\mathcal{T}_j^{(i)}|$ denotes the cardinality of the set $\mathcal{T}_j^{(i)}$.

$$\mathbf{E}_j^{(i+1)}(j) = \frac{1}{|\mathcal{T}_j^{(i)}|} \sum_{t \in \mathcal{T}_j^{(i)}} \mathbf{Y}_t(j) \qquad (8)$$

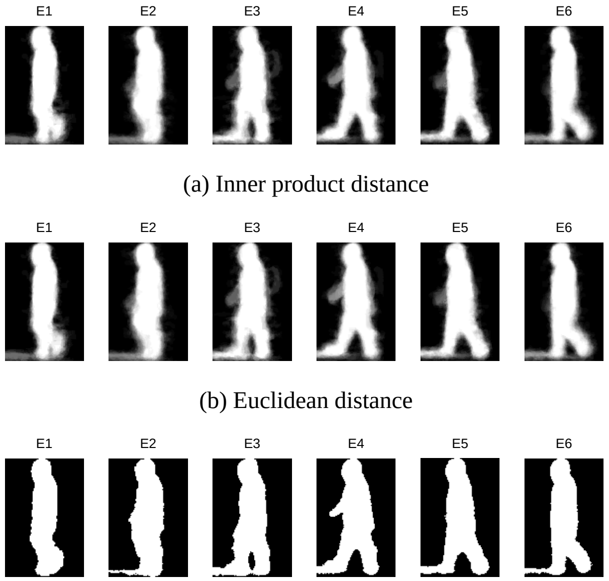
$$\mathbf{E}_j^{(i+1)}(j) = \sum_{t \in \mathcal{T}_j^{(i)}} \tilde{\mathbf{Y}}_t(j) \qquad (9)$$

$$\mathbf{E}_j^{(i+1)}(j) = \operatorname{median}_{t \in \mathcal{T}_j^{(i)}} \{\mathbf{Y}_t(j)\} \qquad (10)$$

The exemplars estimated for one observation sequence using the three distance metrics in (5), (6), and (7) are displayed in Fig. 3. Given $\mathcal{E}^{(i+1)}$ and $A^{(i)}$, we can calculate $A^{(i+1)}$ using the Baum-Welch algorithm [1]. Thus we can iteratively refine our estimates of the HMM parameters. It usually takes only a few iterations in order to obtain an acceptable estimate.

3.3. Identifying from a Test Sequence

Identifying the model index from a sequence involves deciding which of the model parameters to use for discrimination. Given the models in the gallery, $\mathcal{L} = \{\lambda_1, \lambda_2, \ldots, \lambda_P\}$ and the probe sequence, $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T\}$, we would like to find the model and the path that maximises the probability of the path given the probe sequence. The ID is obtained as in (2). We do not need to use the trained parameter set, λ, as a whole. For example, if we believe that the transition matrix is predominantly indicative of the speed at which the subject walks, and is therefore not suitable as a discriminant of the ID of the subject, then we have the option of using only part of the parameter set given by $\gamma_p = (B_p, \pi_p)$ instead of using the HMM parameter set in its entirety. In this case, the conditional probability of the sequence, given the ID, is given as follows. The Baum-Welch algorithm could be used in order to obtain $A_p^{\mathcal{X}}$ recursively in (12).



(a) Inner product distance

(b) Euclidean distance

(c) SAD distance

Fig. 3. Exemplars estimated using various distance measures

$$\Pr[\mathcal{Q}|\mathcal{X}, \gamma_p] = \Pr[\mathcal{Q}|\mathcal{X}, A_p^{\mathcal{X}}, \gamma_p] \qquad (11)$$

$$A_p^{\mathcal{X}} = \arg_A \max \Pr[\mathcal{X}|A, \gamma_p] \qquad (12)$$

4. EXPERIMENTAL RESULTS

The objective of our experiments was to evaluate the performance of the algorithm and also compare the efficacy of the different distance measures in gauging the similarity between two images as far as posture is concerned. The USF database contains video sequences of 75 individuals a subset of whom feature in sequences collected under each of 8 different conditions. The sequences are labelled *Gallery*, *Probe A*, *Probe B*, *Probe C*, *Probe D*, *Probe E*, *Probe F*, and *Probe G*. The number of valid background-subtracted sequences in some probes is less than the number of actual sequences. We trained our parameters using the sequences from the *Gallery* set. In each experiment, we tried to identify the sequences in each of the seven probe sets from the models obtained from the *Gallery* set using the inner product distance measure. The ID was calculated using (2). The experiments were repeated with different distance measures. The results of the experiment using the IP distance measure between feature vectors in the form of Cumulative Match Scores (CMS) plots [11] are in Fig. 4. Table 1 gives a brief summary of the experiments conducted (G and C denote

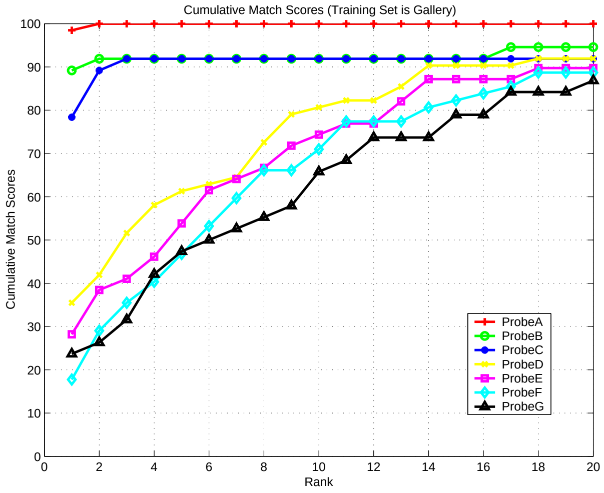


Fig. 4. CMS plots of *Probes A-G* tested against *Gallery*

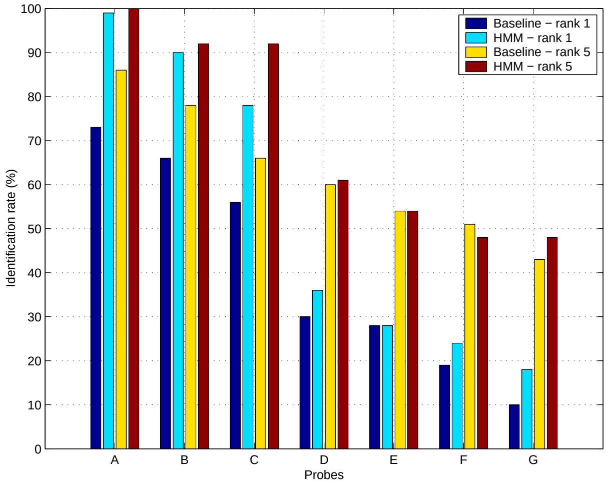


Fig. 5. Comparison of detection rates at ranks 1 and 5

grass and concrete surfaces, A and B denote different shoe types, and R and L denote different camera views). We observe that the distance measure that works best and is most simple to implement is the inner product distance. The performance comparison with the baseline [11] is illustrated in Fig. 5.

5. CONCLUSION

In this paper, we have proposed a general HMM-based framework to represent and recognize human gait. The framework provides algorithms to train the HMM paramters and to identify probe sequences. The framework has the potential to work with suitably complex feature vectors and distance measures that are less susceptible to viewing angles or other factors, though we have used sequences with a constant viewing angle. The periodic nature of gait was exploited in a linear filtering network to obtain good initial values for the exemplars. We have analysed different distance metrics and find that the inner product distance works best. The results obtained are significantly better than the baseline with the additional advantage of compact representation as compared to the baseline. In future work we will study the effect of the background subtraction and other factors such as footwear in order to obtain a more robust and compact feature vector.

Probe	P_I (at rank 1)			P_I (at rank 5)		
	IP	Euclid	SAD	IP	Euclid	SAD
A (GAL) [66]	99%	99%	98%	100%	100%	100%
B (GBR) [37]	89%	89%	89%	92%	92%	92%
C (GBL) [37]	78%	78%	75%	92%	92%	92%
D (CAR) [62]	36%	29%	23%	62%	60%	59%
E (CBR) [39]	29%	28%	21%	54%	54%	59%
F (CAL) [62]	24%	19%	16%	47%	46%	44%
G (CBL) [38]	18%	14%	15%	48%	48%	45%

Table 1. Performance across distance metrics. The numbers in square brackets denote the number of individuals in that set.

6. REFERENCES

[1] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–285, February 1989.

[2] A.F. Bobick and J.W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 23, pp. 257–267, 2001.

[3] P.S. Huang, C.J. Harris, and M.S. Nixon, "Recognizing humans by gait via parametric canonical space," *Artificial Intelligence in Engineering*, vol. 13, no. 4, pp. 359–366, 1999.

[4] C. Bregler, "Learning and recognising human dynamics in video sequences," *Proc. of the IEEE Conf. on Comp. Vision and Patt. Recog.*, pp. 568–574, 1997.

[5] T. Starner, J. Weaver, and A. Pentland, "Real-time american sign language recognition from video using hmms," *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 12, no. 8, pp. 1371–1375, 1998.

[6] J. Yamato, J. Ohya, and L. Ishii, "Recognizing human action in time-sequential images using hidden markov model," *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 624–630, 1995.

[7] B. Frey and N. Jojic, "Learning graphical models of images, videos and their spatial transformations," *Proc. of the Conf. on Uncertainty in Artificial Intelligence*, 2000.

[8] K. Toyama and A. Blake, "Probabilistic tracking in a metric space," *Proc. of the Int. Conf. on Computer Vision*, 2001.

[9] A. Kale, N. Cuntoor, and R. Chellappa, "A framework for activity-specific human recognition," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing*, May 2002.

[10] P.J. Phillips, S. Sarkar, I. Robledo, P. Grother, and K. Bowyer, "The gait identification challenge problem: Data sets and baseline algorithm," *Proc. of the Int. Conf. on Pattern Recognition*, 2002.

[11] P.J. Phillips, S. Sarkar, I. Robledo, P. Grother, and K. Bowyer, "Baseline results for the challenge problem of human id using gait analysis," *Proc. of the 5th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2002.