A FRAMEWORK FOR ACTIVITY-SPECIFIC HUMAN IDENTIFICATION

*Amit Kale, Naresh Cuntoor and Rama Chellappa**

Center for Automation Research
University of Maryland at College Park
College Park MD 20742 USA

ABSTRACT

In this paper we propose a view based approach to recognize humans when engaged in some activity. The width of the outer contour of the binarized silhouette of a walking person is chosen as the image feature. A set of exemplars that occur during an activity cycle is chosen for each individual. Using these exemplars a lower dimensional *Frame to Exemplar Distance (FED)* vector is generated. A continuous HMM is trained using several such FED vector sequences. This methodology serves to compactly capture structural and dynamic features that are unique to an individual. The statistical nature of the HMM renders overall robustness to representation and recognition. Human identification performance of the proposed scheme is found to be quite good when tested on outdoor video sequences collected using surveillance cameras.

1. INTRODUCTION

The need for automated person identification is growing in many applications such as surveillance, access control and smart interfaces. It is well-known that biometrics can be a powerful cue for reliable automated person identification. Established biometric-based identification techniques include fingerprint and hand geometry methods, speaker identification, face recognition and iris identification. However, the applicability of all these methodologies is usually restricted to very controlled environments. For example, current face recognition technology is capable of recognizing only frontal or nearly frontal faces and speaker identification requires the subject to speak sufficiently long. When the problem of person identification is attempted in natural settings, such as those that occur in the automatic surveillance of people in strategic areas, it takes on a new dimension. Biometrics such as fingerprint or iris are then no longer applicable. It is interesting to explore biometric signatures which can be obtained non-invasively from a distance. We explore the possibility of using commonly performed activities e.g. a person walking, a person carrying a package while walking or a person running etc. as a biometric signature. For instance, we know from experience that people often recognize others by simply observing the way they walk. This raises the question of whether body dynamics are sufficiently distinct across humans. Note that like speaker identification the focus here is on characterizing the identity of the person performing the activity as compared to the related problem of activity recognition [1, 2, 3].

In this paper we propose a view based method to characterize the way an individual performs a specific activity. We note that activity has two components : a structural component that comprises factors like height, girth of a person, stride length etc and a dynamical component e.g. the way a person swings his arms,legs etc.. Our approach attempts to integrate these two components for representation and identification. During the course of an activity we identify certain key frames or *exemplars* from the sequence that capture the structural information. Exemplars have been used for visual tracking in Toyama and Blake [4] and video summarization in Frey and Jojic [5]. Our use of exemplars differs from these applications in that our primary goal is recognition. Using the exemplars we map each frame in the video sequence to a lower dimensional *Frame to Exemplar distance (FED)* vector. Using many such FED vector sequences for the individual, we train a continuous density hidden Markov model which captures the dynamic information. The exemplars together with the HMM jointly serve as a succinct representation of the individual. In order to identify an unknown person we obtain his/her FEDs corresponding to each subject in the database and compute the likelihood that the FEDs were generated by the HMMs in the database. The approach has been tested using a few commonly performed activities and the recognition rates have been found to be quite good.

2. PROPOSED METHODOLOGY

We assume that the camera is stationary and that only one person is present in the field of view. Using background subtraction [6] we obtain the silhouettes of the person. The left

*Supported by the DARPA/ONR grant N00014-00-1-0908.

and right boundaries of the body are traced by examining the pixel intensities with a weighted low pass filter from leftmost and rightmost ends of the image. The width of the silhouette along each row of the image is then stored after outliers have been rejected. The width along a given row is simply the difference in the locations of rightmost and leftmost boundary pixels in that row. This "width" feature has the advantage of ease of representation and low computational cost.

Let $\mathcal{X}^{A_i,j} = \{x_1^{A_i,j}, x_2^{A_i,j}, \cdots x_T^{A_i,j}\}$ denote the sequence of width vectors corresponding to the activity A_i where $i = 1, \cdots, Q$ for the person j where $j = 1, \cdots, P$ and T denotes the (variable) length of the activity cycle. We drop the superscript A_i henceforth for the sake of brevity. Using $\mathcal{X}^{A_i,j}$, we wish to build a model which can be used for the task of recognition. Typically x_t belongs to a high dimensional space and shows strong temporal dependencies and modeling it is a challenging problem.

2.1. Exemplars: The Structural Aspect

One possible solution to the problem of representation and recognition of humans from activity lies in a closer examination of the physical process behind the generation of that activity. For example, consider the example of a person walking. During a walk cycle, it is possible to identify certain distinct stances $\mathcal{E} = e_1, \cdots, e_N$. Figure 1 shows the stances for $N = 5$ for two different people. These stances are generic, in the sense that every person transits between these successive stances as he/she walks. They represent the structural component of identity. In general, given the sequence $\mathcal{X}^{A_i,j}$, we wish to pick exemplars which will optimally represent the activity.

For a fixed N we use the K-means clustering algorithm with a weighted mean-squared error distortion measure to design the codebook. The distortion measure is made invariant of translations in the x and y directions. The temporal continuity of the training sequence can be exploited to choose the initial centroids. In order to choose N we computed the average distortion as a function of N and plotted the rate distortion curve (Figure 2). We found that the average distortion does not change appreciably beyond $N = 5$. Hence we choose the number of exemplars N to be 5. Note that it is possible to have other representations of the exemplars. For example, the entire silhouette in each image can be looked upon as an entity and other appropriate distortion measures invariant to specific image transformations can be used. Examples include the Hausdorff distance, shuffle distance [7] etc. In general the distortion measure can be expressed as :

$$\min_{\alpha_1,\cdots,\alpha_l} d(e_i, T(\alpha_1, \ldots \alpha_l)x_j)$$

where d is a generalized distortion measure. $\alpha_1, \cdots, \alpha_l$ represent a set of invariants and T represents a transformation matrix which depends on the α_i's e.g. in the case of Euclidean similarity $l = 3$ $\alpha_1 = u$, where u is the offset, $\alpha_2 = \theta$ where θ is the angle of rotation and $\alpha_3 = s$ where s is the scaling and

$$T(\alpha_1, \alpha_2, \alpha_3)x_j = \alpha_1 + R(\alpha_2)\alpha_3 x_i$$

2.2. HMM: The Dynamic Aspect

The exemplars that we extract from the activity sequence can themselves be used in a naive way for recognition using the quantization error of an incoming frame as a measure of nearness to an element in the database. However such an approach will be sensitive to noise in the observations and more importantly to the presence of structurally similar individuals in the database. To improve the discriminability, the dynamics of the data can be exploited. We note that there is a Markovian dependence across exemplars, through a Markov matrix

$$A = [p(e_i(t)|e_j(t-1))] \tag{1}$$

for $i, j \in \{1, \cdots, N\}$ Also we model the observation vector generated when we are "near" a particular exemplar as being generated by a Gaussian density.

$$P(x|e_j) = \mathcal{N}(x; \mu_j, \Sigma_j) \tag{2}$$

The activity cycle (possibly of varying length) can be viewed as a doubly stochastic process in which the hidden process is represented by the transitions across the exemplars, while the observable is the body image generated when near a particular exemplar. An HMM is best suited for describing such a situation [8]. It is our conjecture that these exemplars can be associated with the states of an HMM where the switch from one exemplar to another can be represented by transition probabilities between states. Training involves learning the HMM parameters $\lambda = (A, B, \Pi)$ from the observation sequences. Here A denotes the transition probability matrix, B is the observation probability and Π is the initial probability vector.

One of the important issues in training is learning the observation probability B. As is well known in statistical pattern recognition, the reliability of the estimates of the elements of the B matrix depend on the number of training samples and the dimension of the observation vector (the curse of dimensionality). This issue is addressed as follows:

We compute the (translation-invariant) weighted Euclidean distance between $x(t)$ from $e_i \in \mathcal{E}$ to build a Frame to Exemplar Distance (FED) $f(t)$ which serves as a lower(N-)dimensional representation of the image at time t i.e. we compute

$$f_j^j(t) = ||x^j(t) - e_l^j||_W \tag{3}$$

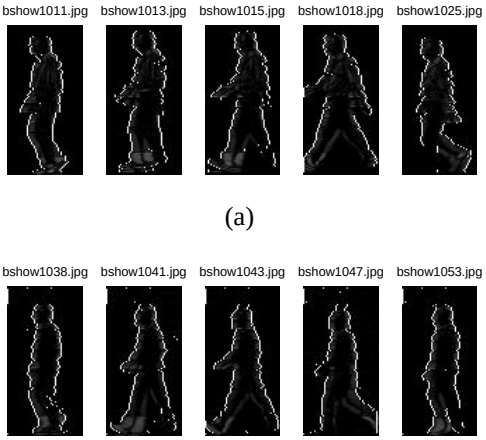

Fig. 1. Five stances corresponding to the gait cycle of a) Person 1 and b) Person 2.

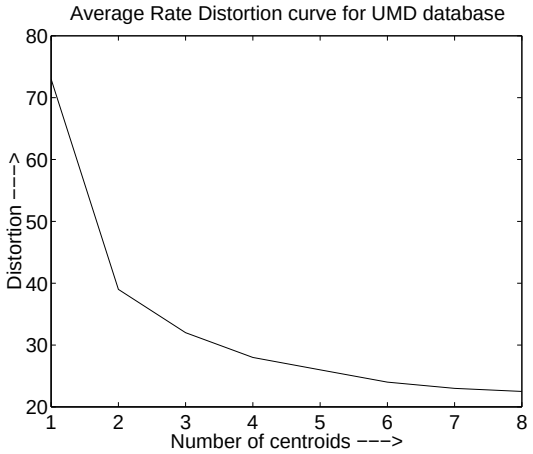

Fig. 2. Rate Distortion curve for number of exemplars vs distortion

where j denotes the person, $t \in \{1, \cdots, T\}$ $l \in \{1, \cdots, N\}$ and W denotes the weighting matrix. The weighting matrix can be used to reflect the relative importance of different regions of the body.

Thus, $f_j^j(t)$ constitutes an observation vector for the HMM model corresponding to person j. Similarly, $f_j^i(t)$ represents the observation sequence of the person i encoded in terms of the exemplars of person j. Note that the dimension of $f_j^j(k)$ is only N. We build the HMM for person j using several such training cycles. The exemplars $\mathcal{E}^j = \{e_1, \cdots, e_N\}$ together with the HMM λ_j jointly serve as the representation for the identity of the person j.

2.3. Recognition

The FED vector has several important characteristics. Firstly, note that by virtue of self-similarity, the encoding of a width vector of person j in terms of the width vectors of the exemplars of person j will yield a lower Euclidean distance than when it is encoded in terms of the width vectors of person i. Thus, loosely speaking, structural information is embedded in the FED vector. Secondly, the manner in which each component of this vector evolves with time encodes the transitional information unique to a person. This transitional information could be the key factor that can distinguish between two individuals who are structurally similar. The overall effectiveness of the FED observation vector in encoding the structural and transitional information is demonstrated in the experimental section.

In order to recognize the unknown person u, the FED vector $f_j^u(t)$ is computed for all $j \in \{1, \cdots, P\}$. We wish to compute the likelihood that the observation sequence f_j^u was generated by the HMM corresponding to the jth person. This can be deciphered by using the forward algorithm which computes this log probability as

$$P_j = \log(P(f_j^u|\lambda_j)) \tag{4}$$

Here λ_j is the HMM model corresponding to the person j. We repeat the above procedure for every person in the database thereby producing $P_j, j \in \{1, \cdots, P\}$. Suppose that the unknown person was actually person m. We would then expect P_m to be the largest among all P_j's , as explained above.

3. EXPERIMENTAL RESULTS

Our experiments were aimed at finding how our methodology performs, with variations in parameters like size of database, speed of performing the activity, clothing and illumination. We considered subjects performing the following activities : normal walk, slow walk on a treadmill, fast walk on a treadmill and walking with an object in hand. We used the following databases

1. UMD database consists of 43 people walking in a T-shaped path. The data was captured by a surveillance camera mounted at a height of approximately 15ft. Each subject has two sequences captured on different days.

2. CMU MoBo database consists of 25 people walking on a treadmill. Each subject has 3 sequences : slow walk, fast walk and walk when carrying a ball.

Training:

Silhouettes of the person performing an activity are extracted using the silhouette extraction procedure described in Section 2. We parse the video sequence in two steps: In the first

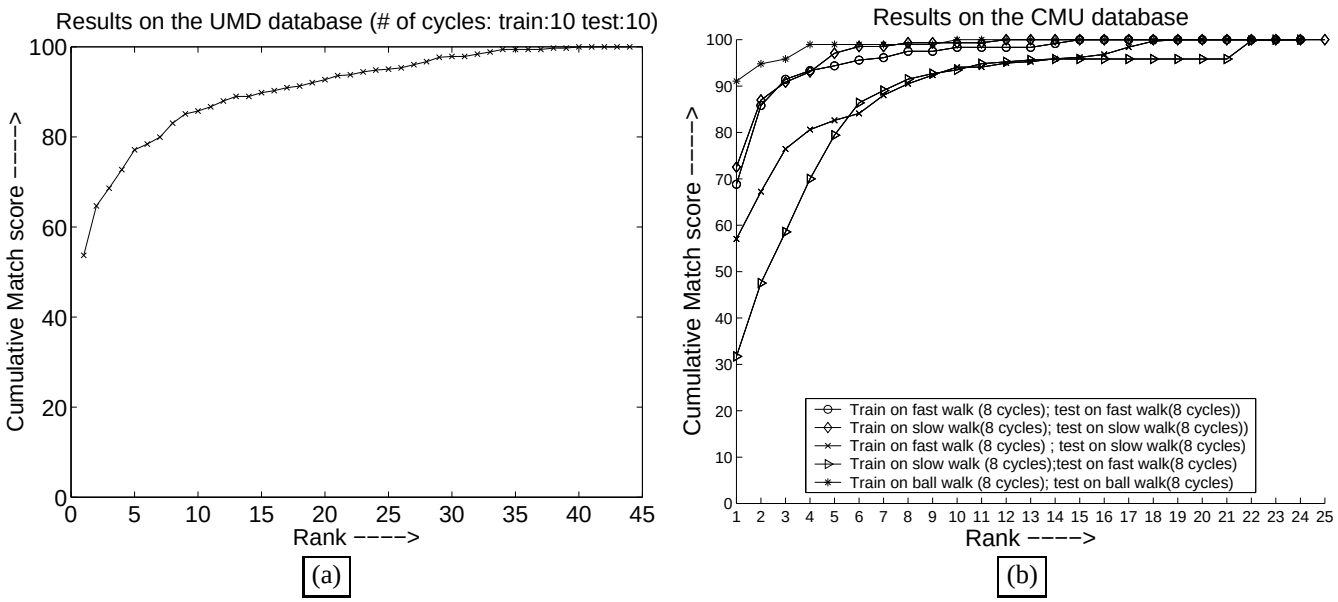

Fig. 3. Identification Performance for (a) UMD database (b) CMU database

parse, we compute the width vector for each frame and identify the start frame of the activity cycle (this is done by identifying local minima of the norms of the width vectors). In the second parse, the FEDs are computed using the exemplars which are identified through k-means clustering. Note that each activity cycle may be of different lengths e.g. during a natural walk a person changes his speed from cycle to cycle. Our methodology is capable of directly handling such variations and does not require an explicit time normalization like [1]. We trained a 5-state, single Gaussian, ergodic HMM λ_i for each person i. For error estimation hold-out method was used.

Recognition:

To perform recognition, given the activity cycles of an unknown person u ; exemplars $\mathcal{E}^j$ and HMMs models λ_j, for $j \in \{1, \cdots P\}$ we generate P sets of FED vector sequences using the width vectors $x^u(t)$ together with each of the e_j,$j \in \{1, \cdots P\}$. For a given FED vector sequence $f_j^u(t)$, we wish to compute the likelihood that model for subject i generated it. For the sake of computational efficiency the Viterbi algorithm is used instead of the forward algorithm to compute P_i in eqn 3. We then rank order the person indices in descending order of the posterior probabilities. Intuitively, the probability P_m will be the largest among $P_1, \cdots, P_N$ if the unknown person u is m, since the patterns of structural and transitional information for u will be the closest to those of m. The procedure is repeated for several activity cycles and the similarity scores are averaged. The graph of rank vs cumulative match scores [9]is plotted . Figure 3 shows the recognition performance for the UMD and CMU databases.

Discussion:

It is natural for a person to change his speed of walking with time. The use of HMM enables us to deal with this variability without explicit time normalization.

We trained the models using the slow walk on the treadmill and tested the fast walk sequences on the treadmill for the CMU database. Observe that the results on CMU database when the HMM is trained using cycles from slow walk and tested using cycles from fast walk, the result is poor compared to the situation when the training and testing scenarios are reversed. This suggests that the model learnt from slow walk is less realistic when trying to model fast walk. In an effort to understand this we ran an experiment whereby we artificially increased the number of frames per activity cycle using interpolation and observed the resulting HMM. It was seen that the A matrix tends towards diagonal dominance. As is well known, this occurs because the basic HMM structure models persistence in a particular state through self loops and hence a geometric distribution. Clearly the geometric distribution does not represent a realistic description of the state duration density in our activity modeling problem. It would be of interest to explicitly model the state duration density as has been done for speech [10]. For the case of training with fast walk and testing on slow walk the dip in performance is caused due to the fact that for some individuals and as biomechanics also suggests there is a considerable change in body dynamics and stride length as a person changes his speed. The result for the UMD database reveals that the performance of the method does not degrade with

an increase in the database size. However the slight drop in performance is due to drastic changes in clothing conditions of some subjects and changes in illumination(causing very noisy binarized silhouettes).

The performance curve for the case where people are carrying a ball in their hands was found to be better than the recognition performance on slow walk and fast walk in the CMU database.

4. CONCLUSION

In this paper, we have proposed a joint exemplar-HMM based approach to represent and recognize humans from activity. A low dimensional observation sequence(FED) is derived from the silhouette of the body during an activity cycle and then a HMM is trained for each person. Human identification is performed by evaluating the log-probability that a given observation sequence was generated by an HMM model.

The method was tested on two different databases. In general, the recognition rates were found to be good. As anticipated, drastic changes in clothing adversely affects recognition performance. The method is sensitive to changes in viewing angle beyond ten degrees. The method is reasonably robust to changes in speed. In the case of human gait recognition we observed in some cases that the stride length changed appreciably with walking speed causing a slight drop in the recognition performance. We observed that the recognition performance for the case of subjects carrying a ball was better than the case of subject walking. This suggests that it might be worthwhile to evaluate the relative significance of the movements of different body parts during an activity.

Presently we are looking at ways to make the scheme invariant to viewing angle and scale which might occur due to the use of multiple cameras. We are also exploring the use of better image metrics to make the FED vector more informative as also appropriate means to model the state duration density . It should be stressed here that the scheme has the potential to distinguish between humans and non-humans. It can also be extended to classify different activities such as walking and running. We are exploring the possibility of activity independent person identification.

5. REFERENCES

[1] 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, no. 3, pp. 257–267, March 2001.

[2] 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, December 1998.

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

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

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

[6] A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," *FRAME-RATE Workshop, IEEE*, 1999.

[7] D. Gavrila and V. Philomin, "Real-time object detection for smart vehicles," *Proc. of the ICCV*, pp. 87–93, 1999.

[8] 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.

[9] P. J. Philips, H. Moon, and S. A. Rizvi, "The feret evaluation methodology for face-recognition algorithms," *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 22, no. 10, pp. 1090–1100, October 2000.

[10] M. Russell and R. K. Moore, "Explicit modelling of state occupancy in hidden markov models for automatic speech recognition," *Proceedings of IEEE Conference on Acoustics Speech and Signal Processing*, June 1985.