COMBINING MULTIPLE EVIDENCES FOR GAIT RECOGNITION

Naresh Cuntoor, Amit Kale and Rama Chellappa*

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

ABSTRACT

In this paper, we systematically analyze different components of human gait, for the purpose of human identification. We investigate dynamic features such as the swing of the hands/legs, the sway of the upper body and static features like height, in both frontal and side views. Both probabilistic and non-probabilistic techniques are used for matching the features. Various combination strategies may be used depending upon the gait features being combined. We discuss three simple rules: the Sum, Product and MIN rules that are relevant to our feature sets. Experiments using four different datasets demonstrate that fusion can be used as an effective strategy in recognition.

1. INTRODUCTION

Biometrics, such as face, voice/speech, iris, fingerprints, gait etc. have come to occupy an increasingly important role in human identification due, primarily, to their universality and uniqueness. Face recognition systems have good performance with canonical views at high resolution and good lighting conditions. Current iris recognition systems are designed to work when the subjects are placed at relatively close distances from the imaging system. A possible alternative is gait or simply, the way a person walks. While medical studies [1] have shown that gait is indeed a unique signature of humans, all the components considered, psychophysical evidence [2] also points to the viability of gait recognition. Gait, a non-intrusive biometric, can be captured by cameras placed at a distance. Illumination changes are not a cause for serious concern. In particular, it might even be attempted in night-time conditions using IR imagery. The potential applications of gait analysis/recognition systems include access control, surveillance and activity monitoring and kinesiology.

We know from our experience that gait and posture provide us with cues to recognize people. Consider a familiar person walking at a sufficiently large distance so that the face is not clearly visible to the naked eye. To recognize the person, we may try to combine information such as posture, arm/leg swing, hip/upper body sway or some unique characteristic of that person. Generally speaking, information may be fused in two ways. The data available may be fused and a decision can be made based on the fused data or each signal/feature can be matched separately, using possibly different techniques and the decisions made may be fused. The former is called data fusion while the latter is decision fusion. Kokar et al. [3] have shown that decision fusion is a special case of data fusion. Note however, that the converse relationship need not be

true. Consequently, data fusion, which tends to be more complex to implement, need not be a bottleneck.

In this paper, we investigate different techniques to combine classification results of multiple measurements extracted from the gait sequences and demonstrate the improvement in recognition performance. Three different sets of features are extracted from the sequence of binarized images of the walking person. Firstly, we investigate the swing in the hands and legs. Since gait is not completely symmetric in that the extent of forward swing of hands and legs is not equal to the extent of the backward swing, we build the left and right projection vectors. To match these time-varying signals, dynamic time warping is employed. Secondly, fusion of leg dynamics and height combines results from dynamic and static sources. A hidden Markov model is used to represent the leg dynamics [4]. While the above two components consider the side view, the third case explores frontal gait. We characterize the performance of the recognition system using the cumulative match scores [5] computed using the aforesaid matrix of similarity scores. As in any recognition system, we would like to obtain the best possible performance in terms of recognition rates. Combination of evidences obtained is not only logical but also statistically meaningful. We show that combining evidence using simple strategies such as Sum, Product and MIN rules improves the overall performance.

The paper is organized as follows: section 2 discusses different features viz. hand and leg swing, leg dynamics, and height, foot dominance and frontal gait. Section 3 presents the experiments performed on different datasets and Section 4 concludes the paper.

2. METHODOLOGY

We assume that, within the field of view of the stationary camera, only one person is present. This simplifies the task of tracking. Background subtraction [6] is used to convert the video sequence into a sequence of binarized images in which a bounding box encapsulates the walking subject. All the features of interest are extracted from the aforesaid sequence of binarized images. Three aspects of gait are discussed: Motion of the hands and legs, dynamics of the legs alone and frontal gait. We address the issue of foot dominance as well. Different strategies such as Sum, Product and MIN rules [7], as applicable in each of the cases are used.

The *left* and *right projection vectors* are constructed from the image sequence to study the motion of the hands and legs. Dynamic time warping is used to match the two vector sequences separately. The overall similarity score is taken to be the sum of the two scores. Secondly, the truncated *width vector* captures the leg dynamics. Hidden Markov model is used to describe the mo-
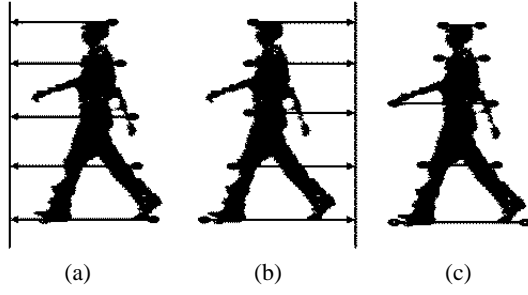
Fig. 1. Illustrating the generation of (a) left projection vector, (b) right projection vector and (b) width vector.

Fig. 2. Effect of eigen decomposition and reconstruction on the width vectors. (a) Overlapped raw width vectors (b) Smoothed width vectors.

tion of the leg within a walk cycle. In the evaluation phase, the absolute value of the forward log probability is recorded as the similarity score. These scores are weighted by a factor that depends on the height of the subject. Thirdly, frontal gait sequences are represented using the width vector, suitably normalized for apparent changes in the height as the subject approaches the camera. A set of width vectors are built for the side view and the two are matched, separately, using DTW. Again, the Sum rule is used to combine the two similarity scores.

2.1. Motion of the arms and legs

In the four-limb system, we seek to find a consistent pattern by systematically analyzing (a) all the four limbs and (b) a pair of limbs. If the degree of coupling between, say, the legs is significantly more than the coupling between the right leg and left hand, then we would assign a higher weight to the similarity score obtained by comparing the leg motion in the reference and test pattern. We first consider the arms and legs of the subject. While it is tempting to assume that gait is a symmetric activity, there exists an asymmetry between the forward and backward swing of the limbs. Maintaining this dichotomy, we build the left and right projection vectors as follows. Given a binarized image, we first align the box so that the subject is in the center of the bounding box. The left and right projection vectors are computed as illustrated in Figure 1 (a) and (b) respectively.

After feature selection and extraction, the next logical step is matching. Direct frame-by-frame matching is not a realistic scheme since humans may slightly alter the speed and style of walking with time. Instead of restricting the frames of possible matches, it would be prudent to allow a search region at each time instant during evaluation. Dynamic Time Warping (DTW) provides for such a mathematical framework [8] in that it allows for non-linear time normalization. We form two matrices of similarity scores by matching the left and right projection vectors in the gallery (reference/training) with those in the probe (testing) set, separately.

The overall similarity score is the sum of the similarity scores obtained the two sets of projection vectors. If the estimation errors of the different classifiers are assumed to be uncorrelated and unbiased, then variance reduces to $\hat{\sigma}^2 = \sigma_e^2/C$.

Like hand dominance (right/left handedness), foot dominance (right/left leggedness) also exists. While matching therefore, we may assume that improperly aligned (i.e. right/left leg forward) reference and test sequences affects the performance. This is an issue because it is not possible to distinguish between the left/right
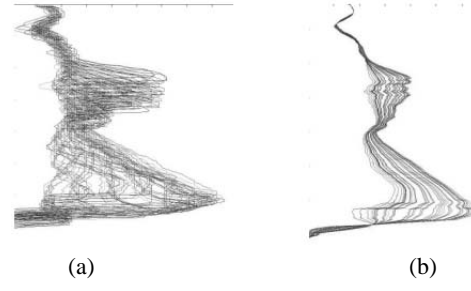
limbs from 2-D binarized silhouettes. Suppose there are five (half-) cycles in both the gallery and probe sequences for a particular subject. To account for foot-dominance, we match the first four half-cycles of the two sequences and generate a matrix of similarity scores. Then, we match the gallery sequence with a phase-shifted probe sequence to generate another matrix of similarity scores. Of the two phase-shifted test sequences, only one can provide a match that is in-phase unless the subject does not exhibit foot dominance. Without loss of generality, we may assume that foot dominance exists in all subjects. Then one of the two test sequences is a better match unless corrupted by noise. Therefore, the two similarity scores are combined using the MIN rule.

2.2. Leg dynamics

Previously, both the hands and legs were considered while selecting the features. If the movement of the hands is restricted (if the subject is carrying an object in his/her hands) or if the sequence is excessively noisy in the torso region due to a systematic failure in background subtraction, then leg dynamics carries information about the subject's gait. We construct a 'width vector' (width of the outer contour of the binarized silhouette) of size $N \times 1$ from each of the images of size $N \times M$ in the sequence, as illustrated in Figure 1(c). Resistance to noise is provided in two stages. While a part of the noise is removed during the computation of the width vector using the spatial correlation of pixels, eigen decomposition and width vector reconstruction utilizes the temporal nature of the data. The sequence of width vectors (matrix of width vectors) $\mathcal{W} = \{W_k, k = 1, 2, \ldots, F\}$ where W_k represents the width vector of size $N \times 1$, at time $t = k$, is standardized and the scatter matrix computed. Eigen decomposition yields the eigen vectors, the largest K of which are retained. The projections of the width vectors on the $K-$ largest eigen vectors yield coefficients that are in turn, used to reconstruct the gait sequence by summing the appropriately weighted $K-$ largest eigen vectors. Figure 2 illustrates the effect of 'eigen-smoothing' on the gait sequence.

A cursory examination of the width vectors suggests that the leg region may exhibit a more consistent pattern compared to other parts of the body such as the arms. At the same time, the gross structure of the body, as contained in the say, the height is also useful in discriminating between subjects. While leg dynamics concentrate on the variation of the width vector in the horizontal direction in the leg region alone, the height of the subject varies in an orthogonal direction. The width vector is truncated so that only the information about the leg is retained. This sequence of trun-
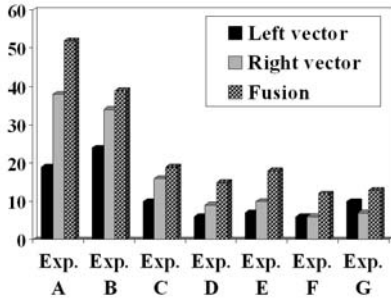
Fig. 3. Identification rates for USF Database: Effect of fusion of left and right projection vectors. Gallery in all the experiments is sequences from surface: grass, shoe type: A, camera view: right.

cated width vectors is the first feature set, say set $\mathcal{A}$. We estimate the height of the subject from the image sequence using robust statistics. The estimated height of the individuals forms the second feature set, say set $\mathcal{B}$. Euclidean distance is used to compare the feature set $\mathcal{B}$ of estimated height of the subjects in the probe and gallery sets.

To compare the truncated width vectors that contain the information about leg dynamics, we use the Hidden Markov model (HMM) [8], which is a generalization of the DTW framework. There exists a Markovian dependence between frames since the way humans go through the motion of walking has limited degrees of freedom. K-means clustering is used to identify 'key frames' or 'stances' during a half-cycle. We found that a choice of $k = 5$ is justified by the rate-distortion curve. We project the sequence of images on the stance set creating a $5 - D$ vector (Frame-to-Stance Distance or *FSD*) representation for each frame and use these samples to train an HMM model using the Baum-Welch algorithm [9]. The viterbi algorithm is used in the evaluation phase to compute the forward probabilities. The absolute values of the log probability values are recorded as the similarity scores.

If the decisions made are statistically independent, we may write the final error probability $\mathcal{P}_e = \prod_{c=1}^{C} \mathcal{P}_e^c$. In practice, however it is difficult to validate this assumption. Instead, we use the low correlation of decisions across feature sets as corroboration to the hypothesis that the errors in the two feature sets, the leg dynamics and the height, are uncorrelated. We use the product rule to combine the scores to compute the overall similarity scores.

2.3. Frontal gait

Hitherto, we have studied gait in its canonical view so that the apparent motion of the walking subject is maximal. This does not preclude the possibility of using other views ranging from the frontal view to any arbitrary angle of viewing. Even in the frontal view where the apparent leg/arm swing is the least, there may be several cues that can be used toward human recognition. More specifically, the head posture, hip sway, oscillating motion of the upper body among other features may pave the way for recognition. As before, to focus our attention on gait, we extract the outer contour of the subject from the binarized gait sequence in the form of the width vector, suitably normalized for an apparent change in height as the subject approaches the stationary camera.

For matching these sequences, we use the DTW technique for similar reasons as outlined in section 2.1. When both the frontal

Table 1. Cumulative match scores at rank 1 and rank 5 for CMU dataset: Combining leg dynamics and height using Sum rule

Feature	CMS at rank 1	CMS at rank 5
Leg dynamics	91	100
Fusion: leg dynamics and height	96	100

Table 2. Cumulative match scores at rank 1 and rank 5 for CMU dataset: effect of frontal and side gait fusion

Feature	CMS at rank 1	CMS at rank 5
Frontal Gait	91	95
Side gait	93	95
Frontal and side	96	97

and fronto-parallel (side) gait sequences are available, it is natural to combine these two orthogonal views before making the final decision about the identity of the subject. One way to combine multiple views is through the use of 3-D models. Currently, 3-D models have been built using sequences captured inside the lab under controlled conditions.[10] takes the visual hull approach while Bobick et al. extract parameters insensitive to the angle of viewing [11]. We adopt the decision fusion approach and combine the matching scores obtained by matching the frontal and side gait sequences separately using the Sum rule.

3. EXPERIMENTS

We report our experiments using the following datasets.

- CMU Dataset (http://hid.ri.cmu/edu)
  consists of 25 subjects walking on a treadmill. Seven cameras are mounted at different angles and we use two of the views for our experiments, viz. the frontal and the side views. The first half of the gait sequence is used for training while the second half is used for testing.

- MIT dataset (http://www.ai.mit.edu/people/llee/HID)
  consists of side view of outdoor gait sequences of 25 subjects collected on four different days. Four experiments are designed. Data from three days provides the training data and data from the fourth day is used as the test sequences.

- UMD dataset (http://degas.umiacs.umd.edu/hid)
  contains outdoor gait sequences captured by two cameras (frontal and side views). 44 subjects are recorded in two sessions. We train with the video data collected from the first session and test with that of the second session.

Table 3. Cumulative match scores at rank 1 and rank 5 for UMD dataset: effect of frontal and side gait fusion

Feature	CMS at rank 1	CMS at rank 5
Frontal Gait	66	86
Side gait	58	74
Frontal and side	85	95

Table 4. Cumulative match scores at rank 1 and rank 5 for UMD dataset: Foot dominance and effect of fusing evidence from two gait sequences (each 4 half cycles long), with one sequence being phase-shifted.

Feature	CMS at rank 1	CMS at rank 5
First sequence	68	84
Phase shifted sequence	70	88
Fusion	77	89

Table 5. USF Dataset: 7 probe sets with the common gallery being G,A,R consisting 71 subjects. The numbers in the brackets are the number of subjects in each probe set.

Experiment	Probe	Difference
A	G,A,L (71)	View
B	G,B,R (41)	Shoe
C	G,B,L (41)	Shoe, View
D	C,A,R (70)	Surface
E	C,B,R (44)	Surface, Shoe
F	C,A,L (70)	Surface, View
G	C,B,L (44)	Surface, Shoe, View

- USF dataset (http://marathon.csee.usf.edu/GaitBaseline/) consists of outdoor gait sequences of 71 subjects walking along an elliptical path on two different surfaces (Grass and Concrete) wearing two different types of footwear (A and B). Two cameras, R and L capture that data. Seven experiments are set up5.

Table 1 shows that while the leg dynamics, by itself has rich information fusion can only improve the performance. Results obtained using the leg dynamics in the cases of UMD and MIT datasets are shown in Tables 6 and 7 respectively. Figure 4 shows that foot dominance is indeed present in certain individuals in the database and that fusing classification results from out of phase gait-sequences serves to increase identification rates. Figure 3 suggests that asymmetry about a vertical axis in the side view may be addressed by considering the two halves of the body on either side of the vertical axis. The results of matching left and the right projection vectors separately were combined using the Sum rule. Tables 2 and 3 show that the performance of frontal gait recognition can be enhanced by using the side view as well.

We observe, in Figure 3 that the right projection vector which captures the forward swing outperforms the left projection vector. This suggests that, in this database, the forward swing of the hands and legs tends has a lesser degree of variability with time (between the gallery and probe sequences). MIT dataset, unlike the other datasets has a low frame rate. Secondly, errors in background subtraction necessitate frame-dropping. This could be a reason for the poor performance.

4. CONCLUSION

Different features that affect gait such as the swing of the hands and legs, the sway in the body as observed in frontal gait, static features like height were systematically analyzed. Starting with dynamic time warping which is a variant of template matching, a more generalized scheme, the HMM was chosen for matching.

Table 6. Cumulative match scores at rank 1 and rank 3 for MIT dataset: Combining leg dynamics and height by adding the similarity scores.

Evaluation Scheme	CMS at rank 1	CMS at rank 3
Day 1 vs. Days 2,3,4	29	50
Day 2 vs. Days 1,3,4	50	100
Day 3 vs. Days 1,2,4	20	54
Day 4 vs. Days 1,2,3	30	52

Table 7. Cumulative match scores at rank 1 and rank 5 for UMD dataset: Combining leg dynamics and height using Sum rule.

Feature	CMS at rank 1	CMS at rank 5
Leg dynamics	31	65
Fusion: leg dynamics and height	49	72

The matrices of similarity scores between the gait sequences in the gallery and probe sets were computed. Sum, Product and MIN rules were used to combine the decisions made using the separate features. As expected, the overall recognition performance improved due to fusion. Experiments were conducted on four different datasets, each dataset presented different types of challenges.

5. REFERENCES

[1] M.P. Murray, A.B. Drought, and R.C. Kory, "Walking patterns of normal men," *Journal of Bone and Joint surgery*, vol. 46-A, no. 2, pp. 335–360, 1964.

[2] J. Cutting and L. Kozlowski, "Recognizing friends by their walk:gait perception without familiarity cues," *Bulletin of the Psychonomic Society*, vol. 9, pp. 353–356, 1977.

[3] M.M. Kokar and J.A. Tomasik, "Data vs. decision fusion in the category theory framework," *FUSION 2001*, 2001.

[4] A. Kale, N. Cuntoor, and R. Chellappa, "A framework for activity-specific human identification," *Proc. of ICASSP*, May 2002.

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

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

[7] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. on PAMI*, pp. 226–239, March 1998.

[8] B. H. Juang, "On the hidden markov model and dynamic time warping for speech recognition - a unified view," *Technical Journal*, vol. 63, pp. 1213–1243, 1984.

[9] 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.

[10] G. Shakhnarovich and T. Darrell, "On probabilistic combination of face and gait cues for identification," *Proc. of the Int. Conf. on Automatic Face and Gesture Recognition*, 2002.

[11] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. on PAMI*, vol. 23, no. 3, pp. 257–267, March 2001.