Feature Fusion of Face and Gait for Human Recognition at a Distance in Video

Xiaoli Zhou and Bir Bhanu
Center for Research in Intelligent Systems
University of California, Riverside, California 92521, USA
{xzhou, bhanu} @vislab.ucr.edu

Abstract

A new video based recognition method is presented to recognize non-cooperating individuals at a distance in video, who expose side views to the camera. Information from two biometric sources, side face and gait, is utilized and integrated at feature level. For face, a high-resolution side face image is constructed from multiple video frames. For gait, Gait Energy Image (GEI), a spatio-temporal compact representation of gait in video, is used to characterize human walking properties. Face features and gait features are obtained separately using Principal Component Analysis (PCA) and Multiple Discriminant Analysis (MDA) combined method from the high-resolution side face image and Gait Energy Image (GEI), respectively. The system is tested on a database of video sequences corresponding to 46 people. The results showed that the integrated face and gait features carry the most discriminating power compared to any individual biometric.

1. Introduction

It has been found to be very challenging to recognize a person from arbitrary views, especially when one is walking at a distance. To obtain optimal performance, a fusion system, which combines face and gait cues from video sequences, is a practical approach to accomplish the task of human recognition. In previous fusion systems, the integration of information from face and gait has been implemented at match score level or at decision level [5] [9].

While fusion at match score or decision level has been extensively studied in the literature, fusion at feature level has drawn more attention in recent years. The fusion at the feature level to combine human faces at multiple views and palmprints is investigated in personal identification by Gao et al. [2]. The fusion at the feature level is discussed in 3 different scenarios for face and hand recognition by Ross et al. [8]. An enhanced Fisher classifier (EFC), which employs the enhanced Fisher linear discriminant model on integrated shape and texture features, is introduced by Liu [7].

The motivation of our work is to demonstrate feature level fusion of face and gait in human recognition at a distance in video. In comparison to the previous work, the contributions of this paper are as follows:

- In this paper, an innovative video based fusion system is proposed, aiming at recognizing non-cooperating individuals at a distance in a single camera scenario. Information from side face and gait is combined. We use the side face instead of frontal face in the system since a side view of face is more likely to be seen at a distance when one exposes the best side view of gait to the camera.
- Face features and gait features, obtained separately using PCA and MDA combined method, are fused at feature level instead of match score level or decision level. Specifically, face features are extracted from high-resolution side face image, which is constructed from multiple low-resolution video frames. Gait features are extracted from Gait Energy Images (GEI), a spatio-temporal compact representation of gait in video.

2. Technical Approach

The overall technical approach is shown in Figure 1.

2.1. Feature Representation

- **Human Body Segmentation:** We use a simple background subtraction method for human body segmentation. We assume that people are the only moving objects in the scene and estimate the human body bounding box from the resulting binary image. Within the human body bounding box, the horizontal alignment is done by centering the upper half silhouette with respect to its horizontal centroid. The size normalization is done by proportionally resizing each silhouette so that all silhouettes have the same height.
- **Gait Frequency and Phase Estimation:** Regular human walking can be considered as cyclic motion where human motion repeats at a stable frequency. Therefore, it is possible to divide the whole gait sequence into cycles. In a normalized binary silhouette sequence, the time series signal of lower half silhouette size from each frame indicates the gait frequency and phase information. We estimate the gait frequency and phase by maximum entropy spectrum estimation [6] from the time series signal .

The 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006

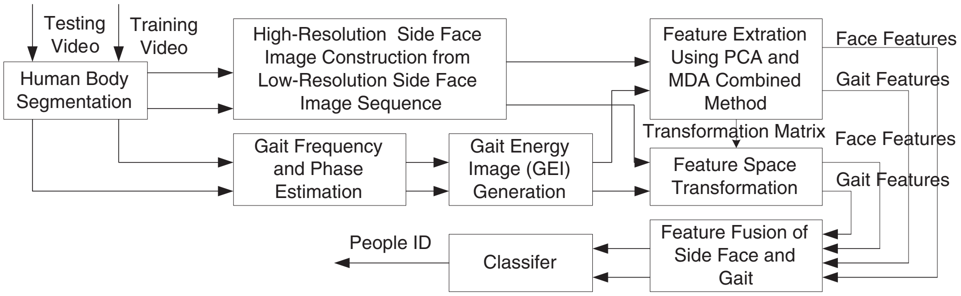
Testing Video
Training Video
Human Body Segmentation
High-Resolution Side Face Image Construction from Low-Resolution Side Face Image Sequence
Gait Frequency and Phase Estimation
Gait Energy Image (GEI) Generation
Feature Extration Using PCA and MDA Combined Method
Face Features
Gait Features
Transformation Matrix
Feature Space Transformation
Face Features
Gait Features
Feature Fusion of Side Face and Gait
Classifer
People ID

Figure 1. Technical approach for integrating side face and gait in video.

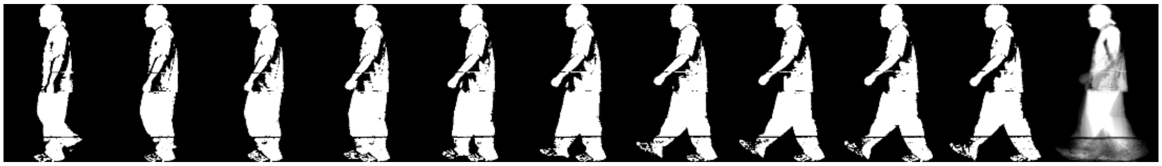

Figure 2. Examples of normalized and aligned silhouette images in a gait cycle. The right most image is the corresponding gait energy image (GEI).

- **Gait Representation:** Given the preprocessed binary gait silhouette image $B_t(x, y)$ at time t in a sequence, the grey-level gait energy image (GEI) is defined as [3]: $G(x, y) = \frac{1}{N}\sum_{t=1}^{N} B_t(x, y)$, where N is the number of frames in the complete cycle(s) of a silhouette sequence, t is the frame number of the sequence (moment of time), and x and y are values in the 2D image coordinate. Figure 2 shows the sample silhouette images in a gait cycle and the right most image is the corresponding GEI. As expected, it reflects major shapes of silhouettes and their changes over the gait cycle.

- **High-Resolution Side Face Image Construction:** Multiframe resolution enhancement seeks to construct a single high-resolution image from multiple low-resolution images. In this work, low-resolution side face images are first localized and extracted by cutting the upper 16% of the extracted human silhouettes obtained from multiple video frames. Then an iterative method [4] is used to construct a high-resolution side face image from the aligned low-resolution side face images. It relies on the fact that the temporally adjacent frames in a video sequence, in which one is walking with a side view to the camera, contain slightly different, but unique, information about face [11]. Figure 3 shows one of the low-resolution face images and the high-resolution reconstructed face image. For comparison, we resize the low-resolution face image using bilinear interpolation.

The normalized side faces are subimages taken from normalized versions of the high-resolution side face images. We obtain normalized side faces based on the locations of nasion, pronasale and throat on the face profile [1]. Examples of normalized side face are shown in Figure 4.

- **Feature Extraction and Feature Space Transformation Using PCA and MDA Combined Method:** In our work, PCA and MDA combined method is used to get low dimensional but effective feature representation for side face and gait, respectively. PCA reduces the dimension of feature space, and MDA identifies the most discriminating features.

Let $\{X_1, X_2, ..., X_K\}$, $X_k \in R^N$, $k = 1, 2, ..., K$, be K random vectors representing K side face images or K GEIs, where N is the dimensionality of the corresponding image. One important property of PCA is its optimal signal reconstruction in the sense of minimum mean square error (MSE) when only a subset of principal components are used to represent the original signal.

$$Y_k = P_{pca}^T X_k \quad k = 1, ..., K. \tag{1}$$

where $P_{pca} = [\Phi_1 \Phi_2 ... \Phi_m], m < N$ and $\Phi_1, \Phi_2, ...\Phi_m$ are orthogonal eigenvectors of the covariance matrix Σ_X. T denotes the transpose operation. The lower dimensional vector $Y_k \in R^m$ captures the most expressive features of the original data X_k.

MDA seeks a transformation matrix W that maximizes the ratio of the between-class scatter matrix S_B to the within-class scatter matrix S_W: $J(W) = \frac{|W^T S_B W|}{|W^T S_W W|}$. $J(W)$ is maximized when the columns of W are the generalized eigenvectors corresponding to the largest eigenvalues of S_W and S_B. c is the number of class and there are no more than $c - 1$ nonzero eigenvalues and the corresponding eigenvectors. The transformed feature vector is obtained as follows:

$$Z_k = P_{mda}^T Y_k = P_{mda}^T P_{pca}^T X_k = M X_k \quad k = 1, ..., K. \tag{2}$$

where $P_{mda} = [\Psi_1 \Psi_2 ... \Psi_n], n < c$ and M is the overall transformation matrix. We can choose n to perform feature selection and dimensionality reduction. The lower dimensional vector $Z_k \in R^n$ captures the most expressive and discriminating features of the original data X_k.

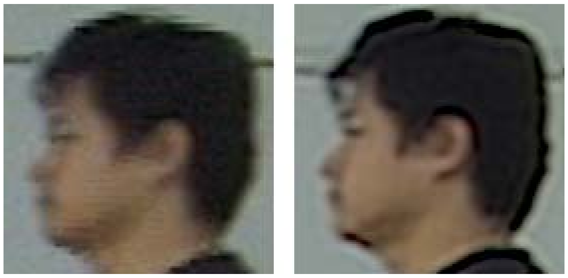

Figure 3. One resized low-resolution face image (left) and one reconstructed high-resolution face image (right).

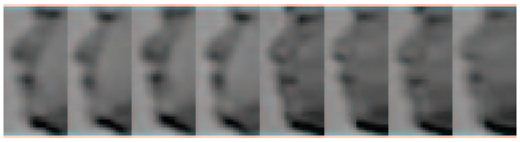

Figure 4. Normalized faces

2.2. Feature Fusion of Face and Gait

• **Feature Normalization Method:** Before feature fusion of face and gait, the individual face features and gait features are normalized to have their values lie within similar ranges. We use a linear method [10], which provides a normalization via the respective estimates of the mean and variance. For N available data of the kth feature, we have

$$\hat{x}_{ik} = \frac{x_{ik} - \bar{x}_k}{\sigma_k} \quad i = 1, 2, ...N \quad k = 1, 2, ...L \tag{3}$$

where $\bar{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_{ik}$ and $\sigma_k^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}(x_{ik} - \bar{x}_k)^2$. L is the number of features. The resulting normalized features have zero mean and unit variance.

• **Classification Method:** Given a side face image F and a gait energy image G, we obtain low dimensional feature vectors, $F' = M^f F$ and $G' = M^g G$, by using PCA and MDA combined method as in Equation (2). M^f and M^g are the overall transformation matrix for face and gait, respectively.

We assume that F' and G' have been normalized before fusion. For feature level fusion, normalized low dimensional features are concatenated to form the new integrated features, which contain both face and gait information. The integrated feature is called *synthetic feature*, defined as

$$H = [F' \; G'] \tag{4}$$

To take advantage of information for a walking person in video, we use all possible combinations of side face features and gait features to generate the maximum number of synthetic feature vectors. Specifically, we have 2 feature vectors of side face and 2 feature vectors of gait for one person from one video. Therefore, we have 4 combinations corresponding to 4 synthetic features for one people from one video. It is reasonable to generate synthetic feature vectors in this way, since the high-resolution face image is built from multiple video frames and GEI is a compact spatio-temporal representation of gait in video.

Let $U_i, i = 1, 2, ...c$, the mean of the training samples of class i after PCA and MDA combined transformation, be the prototype of class i. The unknown person is classified to class k to whom the synthetic feature H is the nearest neighbor.

$$||H - U_k|| = min||H - U_i|| \tag{5}$$

Since we have multiple synthetic features for one person from one video, multiple decisions are obtained using Equation (5). The multiple decisions vote to get the final ID for that person, i.e., the unknown person is classified to the class which gets the maximum votes.

3. Experimental Results

The data is obtained by Sony DCR-VX1000 digital video camera recorder. We collect 92 video sequences of 46 people walking in the outdoor condition and exposing side views to the camera. The video camera operates at about 30 frames per second. The resolution of each frame is 720x480. The distance between people and the video camera is about 10 feet. Each person has two video sequences, one for training and the other for testing. Each video sequence includes one person. Figure 5 shows some examples of the data.

For gait, we can obtain 2 complete walking cycles from a video sequence according to the gait frequency and gait phase. Each walking cycles includes about 20 frames. We construct 2 GEIs corresponding to 2 walking cycles from one video sequence. The resolution of each GEI is 300x200. For face, we also construct 2 high-resolution side face images from one video sequence. Each high-resolution side face image is built from 10 low-resolution side face images extracted from adjacent video frames. The resolution of original low-resolution side face images is 70x70 and the resolution of reconstructed high-resolution side face images is 140x140. After normalization, the resolution of side face is 64x32. Totally, for 46 people, we obtain 92 high-resolution face images and 92 GEIs as the gallery and another 92 high-resolution face images and 92 GEIs as the probe. Correspondingly, we have 92 face feature vectors and 92 gait feature vectors as the gallery and another 92 face feature vectors and 92 gait feature vectors as the probe before fusion. The dimensionality of feature for each side face image is 25 and the dimensionality of feature for each gait energy image is 13.

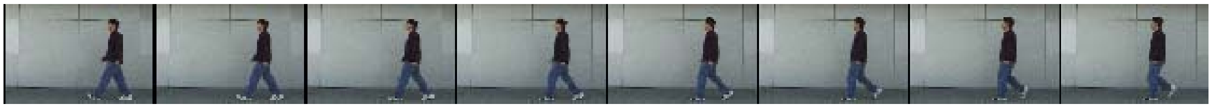

Figure 5. One example of video sequences.

Table 1. Performance at feature level fusion

Fusion Method	LR Face Only	HR Face Only	Gait Only	LR Face and Gait	HR Face and Gait
Recognition Rate	71.7%	84.8%	87.0%	87.0%	91.3%

As explained in Section 2.2, we use 2 face features and 2 gait features of one person from each video to generate 4 synthetic features. These synthetic features take advantage of information in video. The dimensionality of synthetic features is 38, which is derived from 25 features from side face and 13 features from gait. Totally, we have 184 synthetic feature vectors corresponding to 46 people as the gallery and 184 synthetic feature vectors corresponding to 46 people as the probe.

Recognition performance is used to evaluate our method and the quality of extracted features. In our work, the final ID of each person is decided from the voting result using 4 decisions from 4 synthetic features of one person. We also compare the performance using the face features from high-resolution (HR) face images with the performance using the face features from low-resolution (LR) face images. The results are illustrated in Table 1. The individual face features from high-resolution images has better performance at 84.8%, compared with the individual face features from low-resolution images at 71.7%. The individual gait features from GEIs has performance at 87.0%. The integrated features from high-resolution face images and GEIs have the best performance at 91.3%, i.e., 42 out of 46 people are correctly recognized. While the recognition rate using the integrated features from low-resolution images and GEIs is no better than the individual gait features at 87.0%. This proves that using the reconstructed high-resolution face image, we can extract features with more discriminating power. Consequently, we obtain the integrated features with the best discriminating power from face and gait after feature level fusion.

4. Conclusions

In this paper, a video based fusion system is proposed, aiming at recognizing non-cooperating individuals at a distance in a single camera scenario. Information from side face and gait is combined at feature level. The system is tested on a database of video sequences corresponding to 46 people. The best performance is 91.3%, i.e., 42 out of 46 people are correctly recognized, which is achieved using integrated features from high-resolution side face images and GEIs. The results show that our video based recognition method is effective for human recognition in video and the integrated face and gait features carry the most discriminating power than any other individual biometric. In the future, we will mount efforts to overcome some limitations of our approach. For example, we will achieve better alignment of low-resolution side face images and more accurate segmentation of human body from video data.

References

[1] B. Bhanu and X. Zhou. Face recognition from face profile using dynamic time warping. In *17th International Conference on Pattern Recogntion*, volume 4, pages 499–502, 2004.

[2] Y. Gao and M. Maggs. Feature-level fusion in personal identification. In *Proc. of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005.

[3] J. Han and B. Bhanu. Statistical feature fusion for gait-based human recogntion. In *Proc. of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recogntion*, June 2004.

[4] M. Irani and S. Peleg. Motion analysis for image enhancement: Resolution, occlusion and transparency. *Journal of Visual Communication and Image Representation*, 4:324–335, 1993.

[5] A. Kale, A. Roychowdhury, and R. Chellappa. Fusion of gait and face for human identification. In *Proc. Acoustics, Speech, and Signal Processing 2004*, volume 5, pages 901–904, 2004.

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

[7] C. Liu and H. Wechsler. A shape- and texture-based enhanced fisher classifier for face recognition. *IEEE Transactions on Image Processing*, 10(4), 2001.

[8] A. A. Ross and R. Govindarajan. Feature level fusion of hand and face biometrics. In *Proc. of SPIE Conference on Biometric Technology for Human Identification II*, pages 196–204, March 2005.

[9] G. Shakhnarovich and T. Darrell. On probabilistic combination of face and gait cues for identification. In *Proc. Automatic Face and Gesture Recognition 2002*, volume 5, pages 169–174, 2002.

[10] S. Theodorids and K. Koutroumbas. *Pattern recongition*. Academic Press, 1998.

[11] X. Zhou, B. Bhanu, and J. Han. Human recognition at a distance in video by integrating face profile and gait. In *AVBPA*, pages 533–543, 2005.