Q. Wu, L. Wang, X. Geng, M. Li, X. He, 'Dynamic Biometrics Fusion at Feature Level for Video-Based Human Recognition', *Proceedings of Image and Vision Computing New Zealand 2007*, pp. 152–157, Hamilton, New Zealand, December 2007.

Dynamic Biometrics Fusion at Feature Level for Video-Based Human Recognition

Qiang Wu[1], Liang Wang[2],Xin Geng[3], Ming Li[3], Xiangjiang He[1]

[1]Faculty of Information Technology, University of Technology, Sydney, Australia,
[2]Department of Computer Science and Software Engineering, The university of Melbourne, Australia,
[3]School of Engineering and Information Technology, Deakin University, Australia

Email: [1]{wuq, sean}@it.uts.edu.au, [2]lwwang@csse.unimelb.edu.au, [3]{xge, ming.li}@deakin.edu.au

Abstract

This paper proposes a novel human recognition method in video, which combines human face and gait traits using a dynamic multi-modal biometrics fusion scheme. The Fisherface approach is adopted to extract face features, while for gait features, Locality Preserving Projection (LPP) is used to achieve low-dimensional manifold embedding of the temporal silhouette data derived from image sequences. Face and gait features are fused dynamically at feature level based on a distance-driven fusion method. Encouraging experimental results are achieved on the video sequences containing 20 people, which show that dynamically fused features produce a more discriminating power than any individual biometric as well as integrated features built on common static fusion schemes.

Keywords: human recognition, multimodal biometrics, dynamic fusion.

1 Introduction

Gait and face are two cues used frequently for human recognition. The former one is the most suitable biometric for unobtrusive identification at a distance. Most published results of gait recognition were based on human's side view since the quality of motion feature for gait recognition is much higher in the side view than other views. However, it is not always possible to see side view gaits in practical situation. For other views, the motion features required for common gait recognition become less and less notable. On the other hand, face is a widely accepted biometric for human recognition especially when human is viewed from the front. Unlike gait features, face features perform well when people face cameras in a front view. However, the resolution of a human face image will normally go down when the person walks away from the camera. As a consequence, the performance of face recognition decreases.

To obtain better performance, a common solution is to improve the individual algorithm based on a single biometric. In the past decade, there were many different revised methods published for face recognition and gait recognition respectively. Recently, multimodal biometrics fusion [2] has provided another solution since it is more highly efficient than the methods based on one single biometric. There are two broad categories of fusion that are pre-classification fusion and post-classification fusion [3]. Pre-classification fusion refers to combining information prior to the application of any classifier or matching algorithm such as fusions at sensor level and at feature level. In post-classification fusion such as fusions at score level and at decision level, the information is combined after the decisions of the classifiers have been obtained. In terms of fusion rules, static fusion has dominated in multimodal biometrics for a long time [2]. It predefines a fusion scheme to guide the fusion, which does not change during the whole fusion procedure. This paper explores the novel dynamic fusion method which considers the time-dependent changing factors and adjusts the fusion scheme dynamically during the fusion procedure. There have been several approaches reported on fusion of gait and face [4], [5], [6] and [1]. Table 1 summarizes the related work and compares it with the work to be presented in this paper.

The motivation of our work is to demonstrate the capacity of dynamic fusion of face and gait at feature level under the various viewing angles. The proposed algorithm can automatically adjust the fusion scheme without manually selecting high quality feature such as high resolution face image beforehand. People can walk towards cameras, in the side view, or along a direction of the oblique view. In these cases, the distances between cameras and people continuously change. Considering the changing distance, we dynamically integrate static face features with multi-frame based gait features which are obtained directly from the simple binary space-time silhouettes optimized by LPP to produce a more efficient synthetic feature based on multi-model biometrics.

Table 1: Our method for integrating face and gait vs. previous work.

Work	Biometrics	Number of Camera	Data	Fusion Methods
Kale et al [4]	▪ Frontal face ▪ Gait (major side view)	1	▪ 30 subjects ▪ Number of sequence: Not specified	▪ Score level fusion ▪ Sum/Product rule
Shakh-narovich et al [3]	▪ Virtual Frontal face ▪ Virtual Gait (major side view)	4	▪ 26 subjects ▪ 2 to 14 sequences per person	▪ Score level fusion ▪ Min, Max, Sum and Product rules
Zhou et al [1]	▪ Side face ▪ Gait (side view)	1	▪ 46 subjects ▪ 2 sequences per person	▪ Score level fusion ▪ Max, Sum and Product rules
Zhou et al [5]	▪ Side face ▪ Gait (side view)	1	▪ 46 subjects ▪ 2 sequences per person	▪ Feature level fusion ▪ Static fusion
This Paper	▪ Face of various views ▪ Gait of various views	1	▪ 20 subjects ▪ 2 or 4 sequences per person	▪ Feature level fusion ▪ Distance driven dynamic fusion

The paper is organized as follows. The approaches to obtaining gait features and face features respectively are introduced in Section 2. The details of dynamic feature level fusion of face and gait are presented in Section 3. Section 4 demonstrates the experimental results. Conclusions are given in Section 5.

2 Feature Acquisitions of Gait and Face

2.1 Gait Feature Acquisition

Human gait is represented by the temporal variation of human silhouettes. Gait feature will be obtained directly from the silhouette images as shown in [7]. In order to simplify our discussion but without biasing the major concerns of this paper, we assumes that the video to be investigated in this paper has a stable background. Thus, each human silhouette is generated by the common background subtraction methods. Figure 1 shows the example of a silhouette extracted from a video frame.

Each silhouette image is then centred and normalized on the basis of keeping the aspect ratio property of the silhouette so that the resulting images contain as much foreground as possible, do not distort the motion shape, and are of equal dimensions for all input frames. We use the normalized silhouette images as visual inputs to calculate gait features. But it is a formidable task to learn the complete structure of the motion manifold in a high dimensional image space. We use Locality Preserving Projection (LPP) [10] to find low dimensional manifold subspaces for more compact feature extraction and representation.

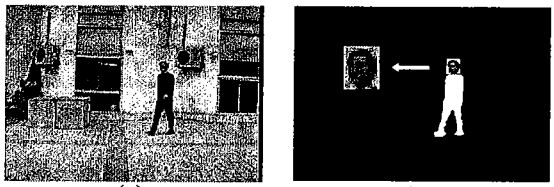
Given P video clips containing gait information of C different people respectively and $M_i (i = 1,2,\cdots,P)$ frames in each video clip, there are totally $N = \sum_i M_i$ frames. The data set is represented by $\Theta = [X_1; X_2; \cdots X_N]^T$ where X_i a row vector is representing a silhouette frame. Θ is regarded as a graph with N nodes. The adjacency graph is constructed based on K-Nearest Neighbours (K-NN) [8]. The link weight between node X_i and X_j is denoted by w_{ij}. Thus, the weight matrix W is a $N \times N$ sparse symmetric matrix. Regarding the assignment of weight, we choose the 0-1 weighting rule. Namely, if there is a link between X_i and X_j according to K-NN, $w_{ij} = 1$. Otherwise, $w_{ij} = 0$. Then, the eigenmaps can be obtained by solving the generalized eigenvector problem,

$$\Theta L \Theta^T e = \lambda \Theta D \Theta^T e \qquad (1)$$

where D ($D_{ii} = \sum_j w_{ji}$ $(j, i = 1,2,\cdots N)$) is a diagonal matrix and. $L = D - W$ is the Laplacian matrix. Let the column vectors $e_1, \cdots, e_l$ be the solutions of Equation (1), which are sorted according to their eigenvalues $\lambda_1 < \lambda_2 < \cdots \lambda_l$. Then, the low dimensional manifold embedding of gait feature is represented by,

$$G = E^T \Theta \qquad (2)$$

where $G = [g_1; g_2; \cdots; g_N]$ and $E = [e_1, e_2, \cdots, e_l]$. So each original silhouette frame feature vector X_i is embedded into a point g_i in the corresponding low dimensional feature space. According to our assumption at the beginning, P video clips containing gait information of C different people respectively and $M_i (i = 1,2,\cdots,P)$ frames in each video clip and



(a) (c)

Figure 1. Silhouette and face extracted from a frame of video: (a) a frame in the video, (b) the extracted face image.

totally $N = M_1 + \cdots + M_P$ frames, the gait feature matrix of video j,

$$\hat{G}_p = [\hat{g}_{p,1}; \hat{g}_{p,2}; \cdots; \hat{g}_{p,m_j}] \quad (3)$$

$\hat{g}_{p,m} = g_t (1 \le p \le P, 1 \le m \le M_p, t = \sum_{p'=1}^{p-1} M_{p'} + m)$ is the feature vector of the m-th frame in the p-th video clip.

2.2 Face Feature Acquisition

After people silhouette is extracted, face detection can be greatly simplified. Based on the observation through extensive empirical experiments, in our experiment data set, the upper 1/6 of the silhouette is located as the face area. An example of the face image is shown in Figure 2 (b). For practical system, mature face detection methods should be used to more precisely locate human face. The extracted face images are first equalized based on histograms. They are then transformed into vectors of unit length to reduce the variation of illumination.

Fisherface [9] is chosen as the face recognition algorithm to find a feature space that maximizes the ratio of the inter-class difference and the intra-class difference by Fisher's Linear Discriminant (FLD) method. Supposing S_B and S_W are the inter-class scatter matrix and intra-class matrix respectively, the optimal projection matrix $R = [r_1, r_2, \cdots r_Q]$ can be calculated by resolving the related eigenvalue problem

$$S_B R = \lambda S_W R. \quad (4)$$

Given K subject classes, the number of nonzero eigenvalues λ will be at most $K-1$. So, in our experiment, the size of the projection matrix $Q = K-1$.

Human face image is extracted in each frame. Supposing each video clip corresponding to a person is represented as a matrix $Y_p (p = 1,2,\cdots P)$ where P is the total number of video clips, each row vector stores the normalized face vector in one frame. That is, $Y_p = [y_{p,1}; y_{p,2}; \cdots y_{p,M_p}]$. $y_{pm} (m = 1,2,\cdots M_p)$ is the face feature vector of the m-th frame in the video p.

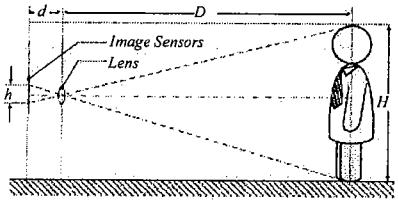


Figure 2. Estimating the distance from the camera to the subject

There are M_p frames in the p-th video clip. Then, after optimization based on Fishface algorithm, the face feature value of a subject will be,

$$F_p = Y_p W \quad (5)$$

where $F_p = [f_{p1}; f_{p2}; \cdots f_{pM_{p_j}}]$ and $p = 1,2,\cdots P$.

3 Dynamic Feature Fusion

To achieve dynamic feature fusion of gait and face, the distance from the camera to the subject in each frame will be estimated. Then, a special weight value will be calculated based on the estimated distance in each frame. This weight value will reflect the different contribution of gait and face in the feature fusion. Finally, a new integrated feature will be constructed. The similarity between the probe video and the gallery video will be measured based on the integrated feature.

3.1 Distance Estimation

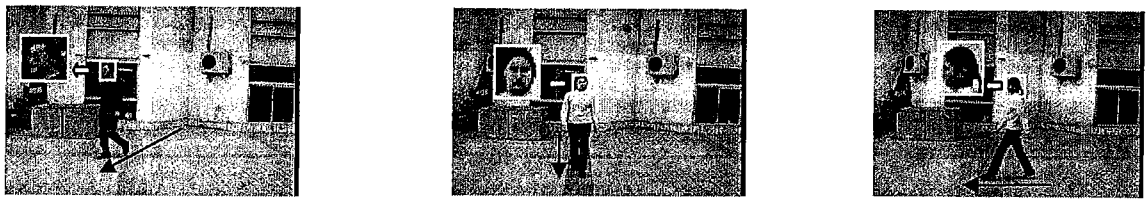
The whole dynamic feature fusion procedure is driven by the distance from subject to camera. In the real situation, the distance between camera and target subject can be measured through additional sensors, stereo image processing techniques, camera calibration techniques and so forth. In this paper, the objective is to testify the capacity of the dynamic feature fusion. Without biasing the major concerns of the paper, we simplify this procedure by estimating the distance based on the basic imaging principle. As shown in Figure 2, H is the actual height of a subject and the corresponding silhouette height on the image is h. Supposing the distance from the subject to the lens is D and the focus of the lens is d, we have,

$$D = Hd/h = \beta/h \quad (6)$$

where $\beta = Hd$. Assuming that video was taken using fixed focus, the variation of h directly reflects the change of the distance from camera to subject.

3.2 Feature Integration

This paper proposes the distance driven weighting concatenation method for feature fusion. Supposing S_{pm} is the new integrated feature vector of



(a) (b) (c)

Figure 3. Typical images in the NLPR Gait Database: (a) the oblique view (45°); (b) the frontal view (90°); (c) the side view (0°)

the frame $m(1 \le m \le M_p)$ in video clip $p(1 \le p \le P)$, we have,

$$S_{pm} = \left[a_{pm} f_{pm}, (1 - a_{pm})\hat{g}_{pm}\right] / \left\| \left[a_{pm} f_{pm}, (1 - a_{pm})\hat{g}_{pm}\right] \right\| \tag{7}$$

$\hat{g}_{pm}$ and f_{pm} were depicted in Equation (3) and (5).

$$a_{pm} = 1/(\gamma D_{pm}{}^2 + 1) = h_{pm}{}^2 / \left(\alpha^2 + h_{pm}{}^2\right) \tag{8}$$

γ is a constant value and $\alpha^2 = \gamma\beta^2$. The construction of function a_{ji} as shown in Equation (8) is based on the fact that the accuracy of face recognition greatly relies on the resolution of face images. Normally, the shorter distance, the higher resolution of the face image has and the more reliable face recognition could be. Moreover, the larger α, the smaller the maximum of a_{pm} is. Suppose $\alpha = h_{\max}$. Then $0 < a_{pm} \le 0.5$. We may adjust the contribution of each individual biometric to the integrated feature by changing α. In order to keep the frame vector value within a regular variation range, the integrated feature vector is normalized by its 2-norm.

3.3 Similarity Measurement

In our experiments, gait feature of each subject is represented as a matrix with multiple feature vectors each of which corresponds to a frame in the video. Accordingly, the new integrated feature is also represented as a feature matrix. Hausdorff distance [7] is used to measure the similarity between gallery video and probe video.

For gait recognition, it is better to include at least one walking cycle in the testing video sequence. In our experiment, each probe video is split into K subsets with overlap along the time axis. Each subset includes enough frames of at least one walking cycle. Let $S_k^{probe} = \left[S_{k1}^{probe}; S_{k2}^{probe}; \cdots; S_{kL}^{probe}\right]$ denote the k-th subset of a probe video clip and there are L frames in each subset. $S_p^{gallery} = \left[S_{p1}^{gallery}; S_{p2}^{gallery}; \cdots; S_{pM_p}^{gallery}\right]$ denotes the integrated feature matrix of the r-th gallery video and there are M_p frames in each video. Gallery videos are not split into subsets. Both S_{kl}^{probe} and $S_{pm}^{gallery}$ can be calculated according to Equation (7). The Hausdorff distance between the probe video and the p-th gallery video (total P videos) is denoted as,

$$d_p = \frac{1}{K}\sum_k \text{Hausdorff}\left(S_k^{probe}, S_r^{gallery}\right), \tag{9}$$

where

$\text{Hausdorff}\left(S_k^{probe}, S_p^{gallery}\right) = \Delta\left(S_k^{probe}, S_p^{gallery}\right) +$
$\Delta\left(S_p^{gallery}, S_k^{probe}\right)$ and

$$\Delta\left(S_k^{probe}, S_p^{gallery}\right) = \text{mean}\left(\min_{1\le l\le L}\left(\left\|S_{kl}^{probe} - S_{pm}^{gallery}\right\|\right)\right).$$

Given the minimum $d_p(1 \le p \le P)$, the unknown person in the probe video will be labeled by $ID(p)$, the subject ID corresponding to gallery video p. That is,

$$Estimated\_PersonID = ID(p),\ p = \arg\min_{1\le p\le P}(d_p) \tag{10}$$

4 Experimental Results

In order to assess the performance of the proposed algorithm, the candidate experimental video data should have the following characteristics: (1) the distance from camera to human continuously changes during the shooting period; (2) the video was taken under the various viewing angles; (3) the quality of both face images and gait video clips is reasonably acceptable during the shooting procedure. After investigating several public database including CMU Mobo Database, CASIA NLPR Gait Database [10], MIT AI Gait Data and Soton Database [11], we selected a subset of NLPR Gait Database. The video clips selected contain subjects walking toward cameras in oblique views (the viewing angle, i.e., the angle between the walking direction and the image plane, is 45°), frontal views (the viewing angle is about 90°) and side views (the viewing angle is about 0°). In total, there are 20 different subjects. There are 2 clips for each subject in the views of 45° and 90° and 4 clips for each subject for the side view of 0°. Some typical images selected from video sequences are shown in Figure 3. During the testing procedure, each probe video is split into a few subsets each of which has 30 frames. The overlap is 15 frames.

Table 2: Recognition Rate Obtained when Using the Difference Subject Features.

View	Single Biometric Features		Integrated Features	
	Gait	Face	Static feature fusion	Dynamic feature fusion
0°	88.8%	53.8%	61.3%	90%
45°	72.5%	75%	85%	90%
90°	72.5%	72.5%	80%	83%

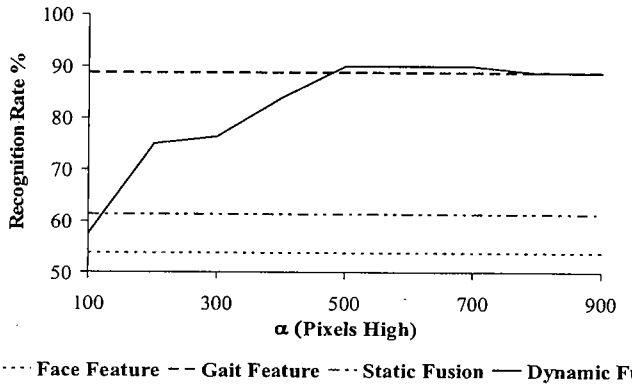


Figure 4. The performance of dynamic fusion under the different weight scheme for 0°

The algorithms were tested based on the Leave-One-Out (LOO) cross validation scheme. Based on LOO, each time only one video clip was selected as the probe video (testing data) and all other clips were used as the gallery videos (training data) . The recognition rates obtained by using different features were recorded, which include single biometric, integrated feature based on static fusion scheme [5], and integrated feature based on distance driven dynamic feature fusion (see Equation 7). For static fusion, a_{ji} in Equation 7 is assigned a constant value, 0.5.

The experimental results are summarized in Table 2. In the views of 45° and 90°, integrated biometric features perform better than single biometric. Normally, gait feature is more reliable in side view (0°) than in other views. On the other hand, face feature performs better only when the major area of face can be seen and the resolution is good enough. In our experiments, the face images in most frames are too small to provide enough resolution, so in the views of 45° and 90° either single face feature or the single gait feature cannot fully perform. However, after they are fused together, their performance will contribute to each other. In terms of integrated feature, the distance driven dynamic feature fusion performs better than the common static feature fusion. The main reason is that static feature fusion does not consider the performance of different biometrics under different conditions, e.g., the distance between camera and human.

For the side view (0°), since it is the best view for the gait recognition algorithm, the single gait feature in this view performs much better than other views. However, since only the side face rather than the major area of face can be seen, the performance of single face feature is bad. The recognition rate is only 54%. In static fusion scheme, both face feature and gait feature contribute equally in the fusion procedure, so the worse face feature seriously affects the relatively better gait feature. Therefore, the integrated feature based on static fusion scheme dose not performs well. It is even worse than the single gait feature. In comparison to the static fusion, dynamic fusion considers the quality of each individual feature. It adjusts the contribution of each feature dynamically during the fusion procedure. In the case of side view video, dynamic fusion scheme includes contribution of gait feature more than face feature, so it performs much better than static fusion and better than any other single feature, i.e., either face or gait. Moreover, from Equation (8), we can see that the value of α determines the maximum value of the weight a_{ji}. In this way, we may adjust the contribution of each individual biometric in the integrated feature by changing α . In our experimental data of side view, the height of silhouette is between 117 and 210 pixels high. By adopting the different values of α , we may obtain the different dynamic fusion performance (see Figure 4.) for the side view video. In Figure 4, we may see that if more face feature is involved in dynamic fusion (smaller α), the performance is not good because the quality of face feature is not good in the side view. If more gait feature contributes to the fusion procedure (larger α), the integrated feature performs much better because the gait feature in the side view is more reliable than other views. When α is larger than a particular value, 800, the

performance of single gait feature and the dynamic integrated feature is almost the same because most contribution while computing the integrated feature (see Equation 7) comes from gait feature.

5 Conclusions

This paper presents a novel dynamic fusion scheme to integrate gait and face biometrics on feature level. In the fusion, since the proposed method considers the different quality of the individual single biometric feature under the different situation (different distance from camera to human in each frame), the performance of human recognition based on the new fusion scheme is not only better than the performance of single biometric but also better than the common static fusion scheme.

The results of this paper indicate that dynamic fusion will be a very encouraging research direction on multimodal biometrics. The basic idea is to assess the quality of individual biometric before fusing them together at different levels such as feature level, score level, and decision level. In this way, we can differentiate the various contributions of the different biometrics based on their performance on different context. Carefully selecting α (see Equation 8) is a key step which will result in the different fusion performance. An adaptive method for selecting better α is being studied. In fact, there are several factors which may influence the quality of individual biometric feature such as imaging distance, viewing angle, illumination and so forth. Successfully finding and understanding these factors for different biometrics will be helpful to construct a reliable synthetic human feature for better human recognition.

6 Acknowledgement

The authors would like to thank the Institute of Automation, Chinese Academic of Sciences (CASIA) for providing the NLPR Gait Database

References

[1] X. Zhou and B. Bhanu, "Integrating Face and Gait for Human Recognition," Proceedings of Computer Vision and Pattern Recognition Workshop, 2006 Conference on, pp. 55, 2006.

[2] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*, vol. 38, pp. 2270-2285, 2005.

[3] C. Sanderson and K. Paliwal, "Information Fusion and Person Verification Using Speech & Face Information," IDIAP, Martigny, Switzerland IDIAP-RR 02-33, 2002.

[4] G. Shakhnarovich and T. Darrell, "On probabilistic combination of face and gait cues for identification," Proceedings of Fifth IEEE International Conference on Automatic Face and Gesture Recognition,, pp. 169-174, 2002.

[5] A. Kale, A. K. Roychowdhury, and R. Chellappa, "Fusion of gait and face for human identification," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04), pp. V-901-4 vol.5, 2004.

[6] X. Zhou and B. Bhanu, "Feature Fusion of Face and Gait for Human Recognition at a Distance in Video," Proceedings of 18th International Conference on Pattern Recognition (ICPR 2006), pp. 529-532, 2006.

[7] L. Wang and D. Suter, "Analyzing Human Movements from Silhouettes Using Manifold Learning," Proceedings of IEEE International Conference on Video and Signal Based Surveillance (AVSS '06), pp. 7, 2006.

[8] X. He and P. Niyogi, "Locality preserving projections," University of Chicago Computer Science, Technical Report TR-2002-09, October 2002.

[9] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 711-720, 1997.

[10] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1505-1518, 2003.

[11] J. D. Shutler, M. G. Grant, M. S. Nixon, and J. N. Carter., "On a large sequence-based human gait database," Proceedings of 4th International Conference on Recent Advances in Soft Computing, Nottingham,UK, pp. 66-71, 2002.