X. Geng, L. Wang, M. Li, Q. Wu, K. Smith-Miles, 'Distance-Driven Fusion of Gait and Face for Human Identification in Video', *Proceedings of Image and Vision Computing New Zealand 2007*, pp. 19–24, Hamilton, New Zealand, December 2007.

Distance-driven Fusion of Gait and Face for Human Identification in Video

Xin Geng[1], Liang Wang[2], Ming Li[1], Qiang Wu[3], and Kate Smith-Miles[1]

[1]School of Engineering and Information Technology,
Deakin University, Victoria 3125, Australia
{xge, ming, katesm}@deakin.edu.au
[2]Department of Computer Science and Software Engineering,
The University of Melbourne, VIC 3010, Australia
lwwang@csse.unimelb.edu.au
[3]Faculty of Information Technology,
University of Technology, Sydney, NSW 2007, Australia
wuq@it.uts.edu.au

Abstract

Gait and face are two important biometrics for human identification. Complementary properties of these two biometrics suggest fusion of them. The relationship between gait and face in the fusion is affected by the subject-to-camera distance. On the one hand, gait is a suitable biometric trait for human recognition at a distance. On the other hand, face recognition is more reliable when the subject is close to the camera. This paper proposes an adaptive fusion method called distance-driven fusion *to combine gait and face for human identification in video. Rather than predefined fixed fusion rules, distance-driven fusion dynamically adjusts its rule according to the subject-to-camera distance in real time. Experimental results show that distance-driven fusion performs better than not only single biometric, but also the conventional static fusion rules including MEAN, PRODUCT, MIN, and MAX.*

Keywords: Human identification, Multi-biometric fusion, Face, Gait, Computer Vision

1 Introduction

Gait and face are two commonly used biometrics for human identification. Both of them are unobtrusive biometric traits, and can be simultaneously obtained by most surveillance systems. The accuracy of most gait recognition algorithms [5] heavily relies on the extraction of motion characteristics. Usually it is easier to recognize the side view gait than the frontal view gait due to the fact that there are more motion characteristics in the side view of a walking person. Up to the present, most reported experiments are performed on the side view gaits. However, it is not realistic to expect only side view gait in real applications. It is interesting to notice that in case of face recognition, the situation happens to be the reverse: there is more information in the frontal face than that in the side face. Thus recognition of the frontal face is generally easier than that of the side face. These complementary properties of gait and face inspires fusion of them to get more accurate results.

Gait is believed to be a suitable biometric trait for human identification at a distance [8]. But the problem of face recognition in such scenario is that the resolution of the face image might be too low to provide enough information for accurate recognition. When the subject is closer to the camera, the resolution increases, and consequently face recognition becomes more reliable. Thus the importance of gait and face in the fusion should dynamically vary according to the distance from the subject to the camera. We call such an approach *distance-driven fusion* of gait and face.

There are several previous works on fusion of gait and face. For example, Shakhnarovich and Darrell [6] proposed to combine virtual gait and face cues generated by a 3D model derived from multiple camera views. Kale *et al.* [4] proposed the fusion of gait and face for a special 'inverted Σ' walking pattern. Zhou and Bhanu [9] proposed a method to improve the side-view gait recognition by using the enhanced side-view face image generated from the video. Table 1 summarizes the main differences between this paper and the previous works. As can be seen, the fusion rules adopted by the previous works are all among the four *static* rules: SUM, PRODUCT, MIN, and MAX, *i.e.*, combine the gait

Table 1: Differences Between This Paper and Previous Works on Gait and Face Fusion.

Work	Biometrics	No. of Cameras	Fusion Rules
Shakhnarovich and Darrell [6]	Virtual frontal face and side gait from a 3D model	4	SUM, PRODUCT, MIN, MAX
Kale *et al.* [4]	Frontal face and 'inverted Σ' gait	1	SUM, PRODUCT
Zhou and Bhanu [9]	Side face and Side Gait	1	SUM, PRODUCT, MAX
This paper	Face and gait in 3 view angles	1	Distance-driven fusion

score and the face score through the operators *sum*, *product*, *min* and *max*, respectively. The rules are predefined and remain fixed when the system is running. Obviously none of them can respond to the changes of the external conditions. In fact, beyond the fusion of gait and face, almost all existing works on multi-biometric fusion are based on fixed fusion rules. On the contrary, the distance-driven fusion is an *adaptive* fusion scheme, which can dynamically adjust the fusion rule according to the external factor that affects the relationship between gait and face in the fusion, *i.e.*, the subject-to-camera distance.

The rest of this paper is organized as follows. The distance-driven fusion of gait and face is proposed in Section 2. The experiments are reported in Section 3. Finally conclusions are drawn in Section 4.

2 Distance-driven Fusion

2.1 Gait Classifier

Human motion can be regarded as temporal variation of human silhouettes. The gait classifier used in this paper is based on the silhouette images [7]. Assume the background to be steady[1], then the silhouette images can be generated through training a Gaussian model for each background pixel over a short period and comparing the background pixel probability to that of a uniform foreground model. One example of the silhouette image is shown in Fig. 1(b). After that, the smallest circumscribing rectangle of each silhouette is centralized and normalized to the same size, and LPP (Locality Preserving Projection) [3] is used to get the corresponding low-dimensional embedding.

In detail, given the training data

$$\mathbf{G} = [\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_n], \quad (1)$$

where $\mathbf{x}_i$ represents the row vector of a binary silhouette image. Assume $\mathbf{G}$ to be a graph with n nodes, an edge will connect nodes i and j if $\mathbf{x}_i$ and $\mathbf{x}_j$ are close. Here 'close' is defined by the K-nearest neighbors. The symmetric $n \times n$ edge matrix $\mathbf{E}$ can be obtained with $g_{ij} = 1$ indicating an edge between nodes i and j, and $g_{ij} = 0$ otherwise. Then the transform matrix $\mathbf{W}_g = [\mathbf{w}_1\ \mathbf{w}_2\ \cdots\ \mathbf{w}_l]$ can be calculated by solving the generalized eigenvector problem

$$\mathbf{GLG}^T\mathbf{w} = \lambda\mathbf{GBG}^T\mathbf{w}, \quad (2)$$

where $\mathbf{B}$ is a diagonal matrix whose entries are column (or row) sums of $\mathbf{E}$, $\mathbf{L} = \mathbf{B} - \mathbf{E}$ is the Laplacian matrix. The $\mathbf{w}_i$'s in $\mathbf{W}_g$ are the solutions of Equation (2) sorted by the corresponding eigenvalues. l is the dimensionality of the subspace. Finally, the projection of a video $\mathbf{X}$ is calculated by $\mathbf{Y} = \mathbf{XW}_g$. Suppose the gallery gait video of person j is $\mathbf{X}_j$, the probe gait video is $\mathbf{X}$, each row ($\mathbf{X}_j(i)$ or $\mathbf{X}(i)$) stores one frame. Then the gait score $G(\mathbf{X}, j) = -d_H(\mathbf{XW}_g, \mathbf{X}_j\mathbf{W}_g)$, where $d_H(m_1, m_2)$ is the mean Hausdorff distance:

$$d_H(m_1, m_2) = \Delta(m_1, m_2) + \Delta(m_2, m_1), \quad (3)$$

$$\Delta(m_1, m_2) = \underset{i}{\text{mean}}(\min_j \|m_1(i) - m_2(j)\|) \quad (4)$$

2.2 Face Classifier

The first step of face recognition is face detection. Since the subject silhouette has already been extracted, face detection can be greatly simplified. Through some empirical experiments, the upper 1/6 of the silhouette is chosen as the face region. One example of the face image extraction is shown in Fig. 1(c). The extracted face images are first histogram equalized and then transformed into a vector of unit length to reduce the variation of illumination.

The face recognition algorithm used in this approach is Fisherface [1], which tries to find a feature space that maximizes the ratio of the inter-person

[1]The proposed method can also be applied to the moving background case, given the proper foreground extraction algorithm, which is out of this paper's scope.

(a) (b) (c)

Figure 1: Silhouette and face extraction: (a) the original image, (b) the extracted silhouette, (c) extraction of the face image.

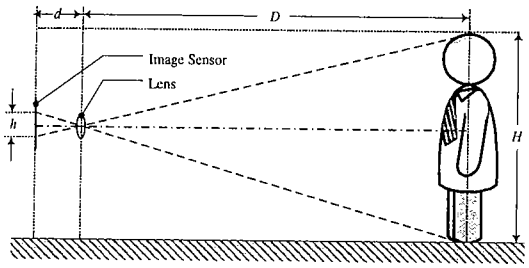


Figure 2: Estimation of the subject-to-camera distance.

difference and the intra-person difference by applying Fisher's Linear Discriminant (FLD). In detail, suppose $\mathbf{\Omega}_B$ is the between-class scatter matrix and $\mathbf{\Omega}_W$ is the within-class scatter matrix, then the projection matrix $\mathbf{W}_f = [\mathbf{w}_1\ \mathbf{w}_2\ \ldots\ \mathbf{w}_q]$ can be calculated by solving a generalized eigenvalue problem

$$\mathbf{\Omega}_B \mathbf{w} = \lambda_i \mathbf{\Omega}_W \mathbf{w}. \quad (5)$$

There are at most $c-1$ nonzero eigenvalues, where c is the number of classes. So in this paper, q is set to $c-1$. Suppose each video is represented as a matrix, each row stores the normalized face vector in one frame, the gallery video of person j is $\mathbf{X}_j$, the probe video is $\mathbf{X}$, then the face score of $\mathbf{X}$ to person j is

$$F(\mathbf{X}, j) = -d_H(\mathbf{X}\mathbf{W}_f, \mathbf{X}_j\mathbf{W}_f), \quad (6)$$

where d_H is the Hausdorff distance defined in Equation (3).

2.3 Fusion Scheme

The whole fusion procedure is driven by the distance from the subject to the camera. A distance can be estimated for each frame in the video. As shown in Fig. 2, suppose the actual height of the subject is H, his/her height in the image is h, the distance from the subject to the lens is D, and the focus of the lens is d, then

$$H/h = D/d, \quad (7)$$

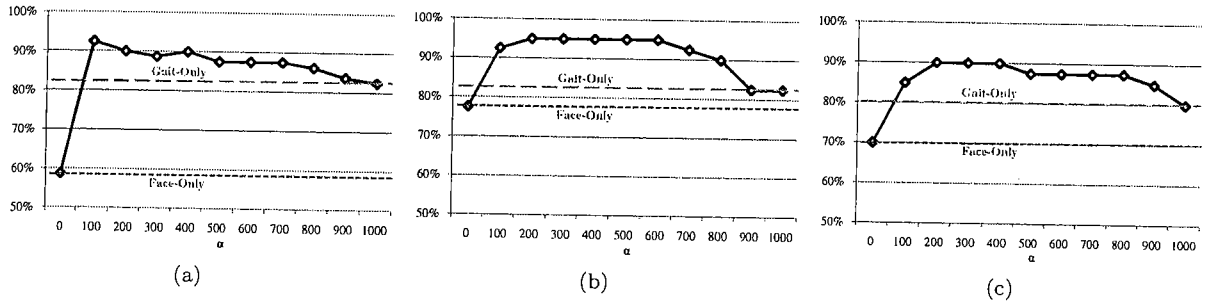
$$D = Hd/h = \alpha/h, \quad (8)$$

where $\alpha = Hd$. Assume that the focus d is fixed and the human height H is approximately the same (when the subject is far away from the camera, D/d is a very large number, the difference in H only has tiny effect on h, thus the normal adult height difference can be ignored), then α is a constant number. h can be calculated from the silhouette image.

As mentioned in Section 1, the reliability of gait and face is affected by the subject-to-camera distance. When the subject is far away from the camera, the resolution of the face image might be too low to be accurately recognized, and thus the fusion system should rely more on gait. When the subject is closer to the camera, the resolution becomes higher, thus face becomes more reliable and deserves more weight in the fusion. A classifier ensemble method called *temporal bagging* is proposed here to dynamically adjust the importance of gait and face in the fusion in real-time. Instead of sampling the data set in the original bagging ensemble [2], the whole video sequence is divided into several subsets (with overlap) along the time axis, each of which corresponds to a period of time. The gait recognition algorithm usually works when the video sequence includes at least one walking cycle (two steps), thus each subset should include at least one walking cycle. The fusion rules in different subsets are different according to the average subject-to-camera distance of each subset. In detail, suppose there are N video clips in the training set $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\}$. Divide each video into m subsets along the time axis, each of which contains p frames including at least one walking cycle. The overlap between the neighboring subsets is v frames. Based on each subset j, a gait classifier (Section 2.1) G^j and a face classifier (Section 2.2) F^j can be trained. Moreover, for each subset j, the average subject-to-camera distance $\bar{D}^j$ can be estimated. Note that although the distance varies within each subset, it is assumed that the variation in approximately one walking cycle is not big enough to apparently change the relationship between gait and face in

Table 2: Recognition Rates on the NLPR Gait Database.

View	Gait-Only	Face-Only	Fusion of Gait and Face				
			Distance-driven	SUM	PRODUCT	MIN	MAX
0°	82.50%	58.75%	**90.00%**	87.50%	87.50%	82.50%	66.25%
45°	82.50%	77.50%	**95.00%**	82.50%	82.50%	90.00%	77.50%
90°	80.00%	70.00%	**90.00%**	82.50%	77.50%	72.50%	87.50%
Avg.	81.67%	68.75%	**91.67%**	84.17%	82.50%	81.67%	77.08%



(a) (b) (c)

Figure 4: Recognition rates of distance-driven fusion with respect to α in (a) the lateral view (0°), (b) the oblique view (45°), and (c) the frontal view (90°).

training. The algorithms are tested in different views (0°, 45° and 90°). For each view, the recognition rates of gait-only, face-only, and the fusion of gait and face are compared. The compared fusion methods include the distance-driven fusion and the fixed fusion rules used by most previous works on multi-biometric fusion [6] [4] [9], *i.e.* SUM, PRODUCT, MIN, and MAX.

3.2 Results

The recognition rates of gait-only, face-only and their fusions in the three different views are tabulated in Table 2. The best performance in each case is highlighted by boldface. The results of the fusion methods that are better than both single biometrics are underlined.

As can be seen from Table 2, in all views, the performance of neither gait-only nor face-only is satisfying. Gait-only performs better at 0° and 45° than 90° because the former two present more motion characteristics. Face-only performs much worse at 0° than the other two views because there is less information in the lateral face than the frontal face. Among the fusion methods, the distance-driven fusion achieves the best performance in all cases, which is remarkably better than those of both single biometrics. As for the static fusion methods, improvement over the single biometric is not guaranteed. In the 0° view, only SUM and PRODUCT achieve better result, in the 45° view, only MIN achieves that, while in the 90° view, only MEAN and MAX can make it. This might be due to the usage of the fixed fusion rules without considering the reliability of different biometrics under different conditions. An unreliable single biometric might deteriorate the performance of the other better biometric in the fusion. Among the four static fusion rules, no remarkable superiority of one over the others can be observed. In summary, when the subject-to-camera distance varies, the relationship between gait and face varies accordingly, thus the dynamic adjustment in the distance-driven fusion is more suitable than the fixed fusion rules.

As mentioned in Section 2.3, the value of α in Equation (8) may also affect the relationship between gait and face in the fusion. The recognition rates of distance-driven fusion with respect to α are shown in Fig. 4. The situations in different view angles are similar. When $\alpha = 0$, distance-driven fusion is the same as face-only. With the increase of α, the recognition rate rapidly achieves the level higher than both face-only and gait-only, and keeps relatively steady until α reaches a high value, say $1,000$, when gait will dominate the fusion so that the recognition rate reduces to the level of gait-only. Note that the relatively broad range of α with steady performance indicates that distance-driven fusion is not sensitive to α, as long as its value is in a reasonable range, say $[100, 800]$.

4 Conclusion

This paper proposes an adaptive fusion method to combine gait and face for human identification in video. Unlike the static fusion rules adopted by most previous works on multi-biometric fusion,

the fusion rule in the distance-driven fusion is dynamically adjusted according to the distance from the subject to the camera in real time. Experimental results show that distance-driven fusion can achieve better performance than conventional fixed fusion rules including MEAN, PRODUCT, MIN, and MAX.

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 7, pp. 711–720, 1997.

[2] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[3] X. He and P. Niyogi, "Locality preserving projections," in *Proc. NIPS'03*, British Columbia, Canada, 2003.

[4] A. Kale, A. Roychowdhury, and R. Chellappa, "Fusion of gait and face for human identification," in *Proc. ICASSP'04*, Montreal, Canada, 2004, pp. V–901–904.

[5] M. S. Nixon, J. N. Carter, M. G. Grant, L. G. Gordon, and J. B. Hayfron-Acquah, "Automatic recognition by gait: progress and prospects," *Sensor Review*, vol. 23, no. 4, pp. 323–331, 2003.

[6] G. Shakhnarovich and T. Darrell, "On probabilistic combination of face and gait cues for identification," in *Proc. FGR'02*, Washington, DC, 2002, pp. 169–174.

[7] L. Wang and D. Suter, "Analysing human movements from silhouettes using manifold learning," in *Proc. AVSS'06*, Sydney, Australia, 2006, p. 7.

[8] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 12, pp. 1505–1518, 2003.

[9] X. Zhou and B. Bhanu, "Integrating face and gait for human recognition," in *Proc. CVPRW'06*, New York City, NY, 2006, p. 55.