Context-Aware Fusion: A Case Study on Fusion of Gait and Face for Human Identification in Video

Xin Geng[a,b,c,*], Kate Smith-Miles[b], Liang Wang[d], Ming Li[e], Qiang Wu[f]

[a]*School of Computer Science and Engineering, Southeast University, Nanjing 210096, China*
[b]*School of Mathematical Sciences, Monash University, VIC 3800, Australia*
[c]*National Key Lab for Novel Software Technology, Nanjing University, Nanjing 210093, China*
[d]*Department of Computer Science, University of Bath, BA2 7AY, United Kingdom*
[e]*School of Information Technology, Deakin University, VIC 3125, Australia*
[f]*School of Computing and Communications, University of Technology, Sydney, NSW 2007, Australia*

Abstract

Most work on multi-biometric fusion is based on static fusion rules. One prominent limitation of static fusion is that it cannot respond to the changes of the environment or the individual users. This paper proposes context-aware multi-biometric fusion, which can dynamically adapt the fusion rules to the real-time context. As a typical application, the context-aware fusion of gait and face for human identification in video is investigated. Two significant context factors that may affect the relationship between gait and face in the fusion are considered, i.e., view angle and subject-to-camera distance. Fusion methods adaptable to these two factors based on either prior knowledge or machine learning are proposed and tested. Experimental results show that the context-aware fusion methods perform significantly better than not only the individual biometric traits, but also

*Corresponding author, Telephone: +86 (25) 5209 0876, Fax: +86 (25) 5209 0876
Email addresses: xgeng@seu.edu.cn (Xin Geng), kate.smith-miles@sci.monash.edu.au (Kate Smith-Miles), lw356@cs.bath.ac.uk (Liang Wang), ming@deakin.edu.au (Ming Li), qiang.wu@uts.edu.au (Qiang Wu)

those widely adopted static fusion rules including SUM, PRODUCT, MIN, and MAX. Moreover, context-aware fusion based on machine learning shows superiority over that based on prior knowledge.
Key words: Multi-biometric fusion, context-awareness, human identification, face recognition, gait recognition.

---

1. Introduction

Biometrics, known as the study of methods for uniquely recognizing humans based upon one or more intrinsic physical or behavioral traits, has become an active research area as well as a widely adopted technique in many real applications. Typical biometric traits include face, gait, fingerprint, iris, voice, vein, hand geometry, etc. Most biometric systems currently deployed in real applications rely on a single source of information, which are called *unimodal biometric* systems. Such systems often suffer from practical problems like noisy sensor data, non-universality and/or lack of distinctiveness of the biometric trait, unacceptable error rates, and spoof attacks [1]. To solve these problems, *multimodal biometric* systems are proposed by combining evidences from different sources [2]. These sources might be multiple sensors [3], multiple classification algorithms [4] or multiple instances [5] for the same biometric trait, or directly from multiple biometric traits [6] [7]. Among them, the systems based on multiple biometric traits, i.e., *multi-biometric fusion*, are generally believed to be more robust than others [8] and thus become the main form of multimodal biometric systems.

Up to the present, most work on multi-biometric fusion is *static fusion*, i.e., the fusion rules are predefined and remain fixed when the system is running. However, in reality, the reliability of a biometric trait might vary with the changes of context.

Table 1: Typical Biometric Traits and Common Influential Context Factors (1 means the context factor may affect the reliability of the biometric, 0 means otherwise)

Biometric Trait			Face	Gait	Fingerprint	Iris	Voice	Vein	Hand Geometry
Context	Human Factors	Pose	1	1	0	0	1	1	0
		Age	1	1	1	1	1	0	1
		Health	1	1	1	1	1	1	1
		Emotion	1	1	1	1	1	0	0
		Occupation	1	1	1	0	1	0	1
	Physical Environment	Location	1	1	0	0	0	0	0
		Noise	0	0	0	0	1	0	0
		Illumination	1	1	0	1	0	1	0
		Background	1	1	0	0	0	0	0
		Humidity	0	0	1	0	0	0	0

Dey [9] defines context as "any information that can be used to characterise the situation of entities". Human factors and physical environment are regarded as the two most important aspects of context [10]. Examples of typical biometric traits and common influential context factors are tabulated in Table 1. As can be seen, the reliability of each biometric trait varies depending on certain context factors. If these traits are to be combined, then the relationship among them in the fusion should accordingly change.

However, static fusion cannot adapt to the changing environment and individual users, which might make multi-biometric systems unstable, unreliable, or even fail to work in real applications. While adaptive information fusion has recently

attracted much attention in several areas, little work has been done for *context-aware multi-biometric fusion*. Some recent work on quality-based multimodal biometrics [11] [12] [13] [14] [15] [16] [17] [18] can be viewed as the first few attempts toward context-aware multi-biometric fusion since the differences in data quality are usually caused by external context factors, such as sensor quality, illumination condition, background noise, etc. In quality-based fusion, a quality assessment algorithm is necessary to calculate a quality score. The assessment usually focuses on the biometric samples themselves, using quality measures directly calculated from the data, such as the signal-to-noise-ratio [14] [11] and the high frequency components of Discrete Cosine Transformation [17]. However, at least at present, a single quality assessment algorithm dealing with all influential context factors is still unrealistic. While we can regard data quality measures as a proxy for knowledge about some external factors, there are certain advantages to trace back to the source context factors of the variability, which include:

1. Quality-based fusion may suffer reluctant responding to new oscillations which are not defined by the quality measures. Adopting the 'divide and conquer' strategy by dealing with the source context factors individually could make the issue much more manageable than trying to pool all variability into a single quality score.
2. Additional devices, such as a laser distance sensor, can be used to detect the variation of context, which are usually much more reliable than quality assessment algorithms.
3. The combination of context factors could be much more complex than a single quality score, such as the context-aware fusion based on neural network, which will be proposed in Section 3.3.2.

Moreover, context means more than those factors determining quality. According to a draft of the INCITS Biometric Sample Quality Standard [19], quality of a biometric sample has three components, namely *character of the source*, *fidelity of the sample to the source*, and *utility of the sample* within a biometric system, which can all be reflected by the accuracy of the biometric system. The basic rule behind quality-based fusion is to give more weight to the more accurate (higher quality) biometric in the fusion. However, a good fusion strategy is not only influenced by the accuracy of individual biometric traits. According to the classifier ensemble theory [20], a good ensemble should be the combination of *accurate* and *diverse* classifiers. Multi-biometric fusion certainly is a special case of classifier ensemble. Thus in addition to *accuracy* (*quality*) of individual biometric, the *diversity* of different biometrics should be considered as well. For example, in the fusion of left-hand fingerprint, right-hand fingerprint and face, usually the former two fingerprint biometrics are more accurate than the face as individual traits. But the similarity between the two fingerprints might prevent them from compensating each other. Considering diversity, the best fusion strategy might be giving higher weights to one fingerprint and the face, while assigning lower weight to the other fingerprint. Since both accuracy and diversity are considered in context-aware fusion while only accuracy is considered in quality-based fusion, the latter can be viewed as part of the former.

To provide more general solutions, a comprehensive framework of context-aware multi-biometric fusion is proposed in this paper. As a typical application, the context-aware fusion of gait and face in video for human identification is investigated. The application scenario is an intelligent identification system deployed in home or workplaces, which can automatically identify the people in a watch list.

Both gait and face are unobtrusive biometric traits and can be simultaneously obtained by most video surveillance systems. The main context factors that affect the relationship between gait and face in the fusion are view angle and distance from the subject to the camera. Usually side view is the best view angle for gait recognition because more motion characteristics can be captured from this angle, while face recognition prefers frontal view because the whole face presents at this angle. Moreover, gait recognition is not very sensitive to subject-to-camera distance (but when subject is too far way, the motion characteristics might partly lose due to low resolution), while face recognition performs better when the subject is close to the camera because face images of higher resolution can be obtained. Thus when view angle or subject-to-camera distance changes, the relative importance of gait and face in the fusion should accordingly change. Methods of incorporating these two context factors into the fusion process with different degrees of freedom, either based on prior knowledge or machine learning, are proposed and compared with conventional static fusion rules, such as SUM, PRODUCT, MIN, and MAX. Note that there are other context factors which might affect the accuracy of gait recognition and face recognition, such as illumination. But the influence of illumination on both biometrics is similar, i.e., both gait and face perform better in good illumination and worse in poor illumination. Thus although illumination variation might affect the accuracy of individual biometrics (and consequently the accuracy of the fusion), it does not apparently change the relationship between them in the fusion. Since the focus of this paper is the influence of context factors on the relationship between modalities in the fusion, the context factors like illumination are not considered in the case of fusion of gait and face.

The rest of this paper is organized as follows. The framework of context-aware

multi-biometric fusion is proposed in Section 2. Then the context-aware fusion of gait and face is investigated in Section 3. Experiments are reported in Section 4. Finally, conclusions are drawn in Section 5.

2. Multi-Biometric Fusion Adaptable to Context

Suppose **B** consists of all information about individual biometric traits used in a multi-biometric system, then static fusion can be expressed by the function

$$\mathbf{F} = f_s(\mathbf{B}), \tag{1}$$

where **F** is the fused information, and the function f_s defines how the components in **B** are combined. The real implements of **B** and **F** depend on the fusion level. For instance, in feature level fusion, **B** consists of all biometric feature vectors of one person, and **F** is the fused feature vector of this person. In score level fusion, **B** consists of the matching scores to each person in the database on each biometric trait, and **F** is the vector of fused matching score to each person. Once f_s is determined, the only factor that can affect **F** is the biometric data **B**. When certain context factors change, the suitable fusion function f_s also changes, but there is no mechanism in static fusion to adapt f_s to the external variations.

To solve this problem, we propose a framework for context-aware multi-biometric fusion, which is shown in Fig. 1. In addition to the biometric data **B**, there are two more inputs into the *Context-Aware Fusion* module. The first is the output of the *Context Monitor* module, i.e., a perceptual signal $\mathbf{M}(t)$ indicating the context at time t. $\mathbf{M}(t)$ is a vector of all context measures. It might come from two different sources: additional sensors (e.g., the devices to detect the atmosphere temperature, humidity, etc.) or the biometric data themselves (e.g., the brightness and contrast

of the images, the emotion and pose of the user, etc.). The other additional input into *Context-Aware Fusion* is the knowledge **K** about the relationship between the fusion rule and the external conditions. It can be prior knowledge from experts or learned from a set of training samples. The real implementation of **K** could be any knowledge representation form, such as a set of rules, a particular function, a neural network, a decision tree, etc. Thus context-aware fusion can be represented by the function

$$\mathbf{F} = f_a(\mathbf{B}, \mathbf{K}(\mathbf{M}(t))). \tag{2}$$

When certain context factors change at time t, $\mathbf{M}(t)$ will capture that, and **K** defines how the fusion rule should be adjusted to adapt to this change. Note that the fusion function f_a is fixed before the system runs. As one parameter of f_a, **K** enables the system to adapt to $\mathbf{M}(t)$ during running time. Thus even if **B** remains the same, the fusion information **F** could be different due to different context factors. Finally, **F** is input into the *Recognition* module to get the user's identity.

During the design of a context-aware multi-biometric system, the following essential questions need to be answered according to specific applications:

1. Which biometric traits are to be combined (determine **B**)?
2. Which context variations might trigger the self-adjustment of the system (determine $\mathbf{M}(t)$)?
3. How does the system respond to the context changes (determine **K**)?
4. How to generate the fused information (determine f_a)?

Among these four questions, (1) and (4) are common for any multi-biometric fusion systems, while (2) and (3) are unique for context-aware multi-biometric fusion. The next section will discuss how gait and face can be combined in a way adaptive to view angle and subject-to-camera distance.

3. Context-Aware Fusion of Gait and Face

As a typical application of context-aware multi-biometric fusion, the adaptive fusion of gait and face in video is investigated in this section. Both gait and face can be extracted from the same source, i.e., the video images, without need of additional sensors. This makes it possible to study multi-biometric fusion based on certain databases for gait recognition, rather than the rare multi-biometric databases, or the virtual 'chimeric persons'. The data used in this paper are the Dataset A and B in the CASIA Gait Database [21]. This is mainly because: (a) Although it is a database for research on gait recognition, the resolution of faces in the video images is also reasonable for face recognition. The faces in most other publicly available gait databases, however, are usually too small to be recognized, such as the currently largest 'HumanID Gait Challenge Data Set' [22]; (b) The variable view angles (walking patterns) and subject-to-camera distance in this database help to illustrate the advantages of adaptive fusion over static fusion. Many other gait databases are either without changes in view angle (e.g., there is only side view in the MIT Gait Database [23]) or without changes in subject-to-camera distance (e.g., in the CMU MoBo Database [24], the subjects walk on a treadmill with fixed distance to the camera); (c) The walking patterns in this database (see Fig. 2 for examples) represent typical natural waking modes, while the walking patterns in some other databases are unlikely to appear in reality (e.g., people walk in circle in the 'HumanID Gait Challenge Data Set' [22]).

The fusion of multiple biometric traits could happen at various levels, such as the feature extraction level, the matching score level or the decision level [6] [8] [25]. In practice, fusion at the matching score level is generally preferred due to the ease in accessing and combining matching scores. There are some previous

Table 2: Differences Between This Paper and Previous Works on Score Level Fusion of Gait and Face

Work	Biometrics	Cameras	Fusion Rules
Shakhnarovich and Darrell [26]	Virtual frontal face and side gait from a 3D model	4	SUM, PRODUCT, MIN, MAX
Kale *et al.* [27]	Frontal face and 'inverted Σ' gait	1	SUM, PRODUCT
Zhou and Bhanu [28] [29]	Side face and side gait	1	SUM, PRODUCT, MAX
Liu and Sarkar [30]	Frontal face and gait around an ellipse	2	SUM, weighted SUM
This paper	Face and gait in 5 view angles	1 (Dataset A) 5 (Dataset B)	Adaptive to view angle and subject-to-camera distance

works on score level fusion of gait and face. For example, Shakhnarovich and Darrell [26] proposed to combine virtual gait and face cues generated by a 3D model derived from multiple camera views. Kale et al. [27] proposed the fusion of gait and face for a special 'inverted Σ' walking pattern. Zhou and Bhanu [28] [29] proposed a method to improve the side-view gait recognition by using the enhanced side-view face image generated from the video. Liu and Sarkar [30] proposed to use both face and gait in enhancing human recognition performance at a distance in outdoor conditions. Table 2 summarizes the main differences between these methods and this paper. As can be seen, the fusion rules adopted by all the previous works are among the four static rules: SUM, PRODUCT, MIN,

and MAX, i.e., f_s in Equation (1) is one of the operators *sum*, *product*, *min* and *max*. None of them can respond to the changes of the external conditions, which creates the need to dynamically adjust the fusion rules according to the context.

The context-aware fusion of gait and face proposed in this paper is also at the score level. As mentioned in Section 1, the two significant context factors that affect the relationship between gait and face in the fusion are view angle and distance from the subject to the camera. Fig. 2 shows the five representative walking patterns in the Dataset A of the CASIA Gait Database. The start and end frame images of the typical video clip for each walking pattern are shown above the corresponding figures. For convenience of description, here we assume that the weighted sum is used to combine the gait score and the face score. The weight assigned to each biometric trait indicates the importance of it in the fusion. Examples of how the gait weight and the face weight vary depending on the context are shown in the figure. Generally speaking, the performance of gait recognition is mainly affected by view angle and not closely related to distance. Usually it will get the best result in the side view (iii) because more motion characteristics can be captured from this angle. As such, the oblique view (ii, iv) is worse and the frontal/back view (i, v) is the worst. On the other hand, face recognition is affected by view angle as well as image resolution, which is determined by the subject-to-camera distance. In contrast to gait, the frontal view (v) is the best angle for face, the oblique frontal view (iv) is the next, then the side view (iii), and then the oblique back view (ii), finally, the back view (i) cannot be recognized at all. Moreover, the closer the face to the camera, the higher the resolution, and the more accurate the recognition. Accordingly, the weights for gait and face in the fusion are adjusted in real-time.

3.1. Matching Scores (Determine **B***)*

First of all, both gait and face need to be extracted from the video images. Fig. 3 illustrates the extraction process. The gait trait is regarded as temporal variation of human silhouettes. 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 extracted binary silhouette image is shown in Fig. 3(c). After that, each silhouette image is centralized and normalized to the same size as in [31]. A gait trait consists of a sequence of such binary silhouette images. While face detection is generally known as a challenging task and many sophisticated detection algorithms have been proposed [32], it can be greatly simplified based on the already extracted silhouette images. According to an anatomical study [33], the head/body length ratio is almost the same for all adult humans, i.e., 0.130. Considering the effects of hairstyle and possible inaccuracy of the silhouette extraction algorithm, here a slightly larger head/body ratio is used, i.e., the upper $1/7$ of the body silhouette is chosen as the face region. One example of the face image extraction is shown in Fig. 3(d). The extracted face images are resized to the same resolution, histogram equalized and then transformed into a vector of zero mean and unit norm to reduce the variation of illumination. Note that in a unimodal biometric system based on face, usually more accurate face detection is required, such as finding the exact locations of the eyes and nose. But when the whole human body is included in the video frames, the face resolution is usually too small for accurate detection of salient facial fea-

[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.

tures (e.g., eyes and nose). Thus here the relatively simple face extraction method is adopted.

3.1.1. Gait Matching Score

Given the training gait data $\mathbf{G} = [\boldsymbol{x}_1^g; \boldsymbol{x}_2^g; \ldots; \boldsymbol{x}_n^g]$, where $\boldsymbol{x}_i^g$ represents the d-dimensional row vector of a normalized binary silhouette image, LPP (Locality Preserving Projection) [34] is used to extract the corresponding low-dimensional features, i.e., find a $d \times l$ ($l < d$) transform matrix $\mathbf{W}_g$ to project a silhouette image into a l-dimensional feature vector $\boldsymbol{y}_i^g = \boldsymbol{x}_i^g \mathbf{W}_g$. Assume the row vectors of $\mathbf{G}$ to be n nodes of a graph, an edge will connect nodes i and j if $\boldsymbol{x}_i^g$ and $\boldsymbol{x}_j^g$ are close. Here 'close' is defined by the k-nearest neighbors. A symmetric $n \times n$ edge matrix $\mathbf{E} = [e_{ij}]$ can be obtained with $e_{ij} = 1$ indicating an edge between nodes i and j exists, and $e_{ij} = 0$ otherwise. Then the transform matrix $\mathbf{W}_g = [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_l]$ can be calculated by solving the generalized eigenvector problem

$$\mathbf{G}^T \mathbf{L} \mathbf{G} \boldsymbol{w} = \eta \mathbf{G}^T \mathbf{A} \mathbf{G} \boldsymbol{w}, \tag{3}$$

where $\mathbf{A}$ is a diagonal matrix whose entries are column (or row) sums of $\mathbf{E}$, $\mathbf{L} = \mathbf{A} - \mathbf{E}$ is the Laplacian matrix. The $\boldsymbol{w}_i$'s in $\mathbf{W}_g$ are the eigenvectors of Equation (3) corresponding to the l largest eigenvalues. Suppose a video is represented by $\mathbf{X}$, where each row $\mathbf{X}(i)$ stores one frame. Then, the features of $\mathbf{X}$ is calculated by $\mathbf{Y} = \mathbf{X}\mathbf{W}_g$. Suppose the gallery gait video (the video stored in the database of known persons) of person p is $\mathbf{X}_p$, the probe gait video (the video to be recognized) is $\mathbf{X}$. Then the gait matching score $s_g(\mathbf{X}, \mathbf{X}_p)$ is calculated by

$$s_g(\mathbf{X}, \mathbf{X}_p) = -d_H(\mathbf{X}\mathbf{W}_g, \mathbf{X}_p\mathbf{W}_g), \tag{4}$$

where d_H is the mean Hausdorff distance defined by

$$d_H(\mathbf{X}\mathbf{W}_g, \mathbf{X}_p\mathbf{W}_g) = \Delta(\mathbf{X}\mathbf{W}_g, \mathbf{X}_p\mathbf{W}_g) + \Delta(\mathbf{X}_p\mathbf{W}_g, \mathbf{X}\mathbf{W}_g), \tag{5}$$

$$\Delta\left(\mathbf{X}\mathbf{W}_g, \mathbf{X}_p\mathbf{W}_g\right) = \underset{i}{\text{mean}}(\min_j \|\mathbf{X}(i)\,\mathbf{W}_g - \mathbf{X}_p(j)\mathbf{W}_g\|). \quad (6)$$

3.1.2. Face Matching Score

The face recognition algorithm used in this approach is Fisherface [35], which tries to find a feature space that maximizes the ratio of the inter-personal difference and the intra-personal difference by applying Fisher's Linear Discriminant (FLD). Suppose each video is represented as a matrix, whose rows store the normalized face vectors in the frames, the transformation matrix of Fisherface is $\mathbf{W}_f$, the gallery video of person p is $\mathbf{X}_p$, the probe video is $\mathbf{X}$, then the face matching score between $\mathbf{X}$ and $\mathbf{X}_p$ is

$$s_f(\mathbf{X}, \mathbf{X}_p) = -d_H(\mathbf{X}\mathbf{W}_f, \mathbf{X}_p\mathbf{W}_f), \quad (7)$$

where d_H is the mean Hausdorff distance defined in Equation (5).

After the scores between the probe video $\mathbf{X}$ to all the known persons in the database have been calculated, the gait scores are stored in the vector $\boldsymbol{b}_1$ with each element corresponding to one gait score (s_g), the face scores are stored in $\boldsymbol{b}_2$ with each element corresponding to one face score (s_f), and the biometric data

$$\mathbf{B} = \{\boldsymbol{b}_1, \boldsymbol{b}_2\}. \quad (8)$$

3.2. View Angle and Distance Estimation (Determine $\mathbf{M}(t)$*)*

In case of context-aware fusion of gait and face, the perceptual signal $\mathbf{M}(t)$ in Fig. 1 consists of two kinds of information: the view angle and the subject-to-camera distance, which can be estimated through analyzing the silhouette images.

Each of the five walking patterns in Fig. 2 corresponds to one view angle θ, which is shown in Fig. 4. In order to keep θ within $[-\pi/2, \pi/2]$, the walking pattern (iv) is flipped left to right. Note that left-right flipping changes little of the

Table 3: Silhouette Analysis to Determine the Walking Patterns

Walking Pattern	View Angle	Conditions to Satisfy	
(i)	$-\pi/2$	$h_e - h_s < -\rho_1 f_h$,	$\lvert c_s - c_e \rvert \leq \rho_2 f_w$
(ii)	$-\pi/4$	$h_e - h_s < -\rho_1 f_h$,	$\lvert c_s - c_e \rvert > \rho_2 f_w$
(iii)	0	$\lvert h_e - h_s \rvert \leq \rho_1 f_h$	
(iv)	$\pi/4$	$h_e - h_s > \rho_1 f_h$,	$\lvert c_s - c_e \rvert > \rho_2 f_w$
(v)	$\pi/2$	$h_e - h_s > \rho_1 f_h$,	$\lvert c_s - c_e \rvert \leq \rho_2 f_w$

relationship between gait and face in the fusion due to human body symmetry. In the video clip of a walking person, suppose the first frame is s, the last is e, the width and height of each frame are f_w and f_h, the height of the silhouette in s is h_s, that in e is h_e, the horizontal position of the silhouette centroid in s is c_s, and that in e is c_e. Then the five walking patterns (i-v) can be determined through Table 3, where ρ_1 and ρ_2 are ratio threshold parameters. Note that the left-right flipping will not change the classification rules, which means flipping pattern (iv) does not require a real image flipping operation. There are more sophisticated ways to detect the view angle, but since the walking patterns in the CASIA Gait Database are relatively fixed, these simple rules are effective enough to estimate the walking pattern (view angle) for each video clip, and they can work very quickly to ensure real-time response. Moreover, even in more general cases when the view angle is not multiples of $\pi/4$, Table 3 can classify the view angle into the closest walking pattern.

The second context factor, the distance from the subject to the camera, can be roughly estimated for each frame in the video as shown in Fig. 5. Suppose the actual height of the subject is H, the height of the silhouette in the image is h, the

distance from the subject to the camera lens is D, and the focal length of the lens is d, then

$$D = Hd/h = \alpha/h, \tag{9}$$

where $\alpha = Hd$. Without knowing α, it seems that D cannot be actually calculated. Fortunately, α can be removed as a common factor from the numerator and denominator when applying the Min-Max normalization:

$$\begin{aligned} \tilde{D} &= \frac{D - \min_i(D_i)}{\max_i(D_i) - \min_i(D_i)} \\ &= \frac{\frac{1}{h} - \min_i(\frac{1}{h_i})}{\max_i(\frac{1}{h_i}) - \min_i(\frac{1}{h_i})}. \end{aligned} \tag{10}$$

Thus α does not affect the normalized distance $\tilde{D}$.

Suppose at time t, the view angle (determined by the walking pattern) is $\theta(t)$, and the normalized subject-to-camera distance is $\tilde{D}(t)$, then the perceptual signal at time t is

$$\mathbf{M}(t) = \{\theta(t), \tilde{D}(t)\}. \tag{11}$$

It is noteworthy that the methods used to estimate θ and D can only give a rough estimation. The relatively good performance of the proposed methods in the latter experiments indicates that a rough estimation of θ and D is enough so long as they reveal the basic variations of the relationship between the component biometric traits. While a precise estimation of θ and D might further improve the performance, the selection of a good fusion rule is much more important, which will be discussed in the next section.

3.3. Context-Aware Fusion Rules (Determine $\mathbf{K}$ *and* f_a*)*

As shown in Fig. 1, the knowledge $\mathbf{K}$ about the relationship between the fusion rules and the context could be determined by prior knowledge or machine learning from a set of training data. In this section, both approaches are investigated, first by a series of empirically determined rules based on weighted sum, then by machine learning based on a neural network. It is worthy to mention that although some methods described in this section are designed for the five representative walking patterns shown in Fig. 2, they can be easily extended to more general cases with arbitrary view angles.

3.3.1. Context-Aware Fusion Based on Weighted Sum

Suppose s_g is the gait matching score calculated by Equation (4), and s_f is the face matching score calculated by Equation (7). Note that with the walking pattern already being estimated by Table 3, the matching scores are calculated between the probe video and those gallery videos of the same walking pattern. Both s_g and s_f are first normalized to the same scale through Min-Max normalization, getting $\tilde{s}_g$ and $\tilde{s}_f$. Then the fusion score is calculated as the weighted sum

$$s = \lambda \tilde{s}_g + (1 - \lambda)\tilde{s}_f, \tag{12}$$

where the weight $\lambda \in [0, 1]$. Recall the function in Equation (2), here the fusion function f_a is weighted sum, and the knowledge $\mathbf{K}$ is about how to generate the fusion weigh λ based on the perceptual signal $\mathbf{M}(t)$ calculated by Equation (11):

$$\lambda = \mathbf{K}(\mathbf{M}(t)). \tag{13}$$

For context-aware fusion, λ is related to the walking pattern determined by Table 3 and the normalized subject-to-camera distance $\tilde{D}$ determined by Equation (10).

With different sensitivity to θ and $\tilde{D}$, λ is generated in four different ways, which are described in detail below (*Case 1* to *Case 4*, from constant to sensitive). The corresponding surface plots in the θ-$\tilde{D}$-λ space are shown in Fig. 6.

Case 1: λ is a constant number 0.5. This case is actually using the static fusion rule SUM. Equal importance is assigned to gait and face no matter how θ and $\tilde{D}$ vary. The corresponding plot is show in Fig. 6(a).

Case 2: λ is a constant number for each view angle. In this case, λ changes with θ but not $\tilde{D}$. The suitable values of λ for each view angle is empirically determined. The basic principle is "gait recognition prefers side view while face recognition prefers frontal view". We experimented with a series of configurations (more details in Section 4) and the highest accuracy was achieved when λ equals 1.0, 0.8, 0.7, 0.6, and 0.5 with θ equals $-\pi/2$, $-\pi/4$, 0, $\pi/4$, and $\pi/2$, respectively. The corresponding plot is show in Fig. 6(b).

Case 3: λ varies with $\tilde{D}$ within a certain range determined by θ. Since the gait recognition algorithm cannot work on a single image, in order to incorporate the influence of $\tilde{D}$, each video clip is divided into m (m might be different for different video clips) subsets along the time axis with an overlap of v frames between the neighboring subsets. Each subset corresponds to a short period of time, and the fusion rule within this period is assumed to be steady. The gait recognition algorithm usually works when the video sequence includes at least one walking cycle (two steps), thus the length of each subset r should include at least one walking cycle. Suppose the average value of $\tilde{D}$ over all frames in the subset i ($i = 1, \ldots, m$) is $\bar{D}^i$, then λ is calculated by

$$\lambda = \frac{\bar{D}^i - \min_i(\bar{D}^i)}{\max_i(\bar{D}^i) - \min_i(\bar{D}^i)} \times (max_\theta - min_\theta) + min_\theta, \tag{14}$$

where $[min_\theta, max_\theta]$ is the range of λ determined by the view angle θ, i.e., the view angle determines the possible range of the weight, and the subject-to-camera distance determines how the weight varies in that range. Note the difference of the min/max operators in Equation (10) and (14): those in Equation (10) regard all frame images in the training set while those in Equation (14) regard the m subsets in a video clip. For each subset i, a fusion score s_i can be calculated by Equation (12). The final score for a video clip is the average score over all its subsets, based on which the person in that video clip is recognized. There are two principles when choosing a suitable range of λ for each view angle. Aside from the one used in *Case 2*, the other one is "the closer the subject to the camera, the smaller λ should be (more weight on face and less on gait)". We also performed a series of experiments on different choices of $[min_\theta, max_\theta]$, and the highest accuracy was observed when the λ ranges are $[1, 1]$, $[0.8, 0.9]$, $[0.7, 0.7]$, $[0.6, 0.9]$, and $[0.5, 0.8]$ for the view angles $-\pi/2$, $-\pi/4$, 0, $\pi/4$, and $\pi/2$, respectively. The corresponding plot is show in Fig. 6(c).

Case 4: λ is generated by a function of θ and $\tilde{D}$. In order to incorporate the influence of $\tilde{D}$, similar to *Case 3*, each video is divided into m subsets and the average distance $\bar{D}$ in each subset is calculated. Finding an optimal function of θ and $\bar{D}$ to generate λ is a difficult problem. But there are still some traces to follow. First, λ should decrease with the increase of θ; Second, λ should increase with the increase of $\bar{D}$; Third, $\lambda = 1.0$ when $\theta = -\pi/2$; Fourth, $\lambda \in [0, 1]$ when $\theta \in [-\pi/2, \pi/2]$ and $\bar{D} \in [0, 1]$. After several empirical tests on possible functions, the following hyperbolic tangent functions are chosen to fit this problem.

$$f(\theta) = a_1 \tanh(b_1\theta) + c_1, \tag{15}$$

$$g(\bar{D}) = a_2 \tanh(b_2\bar{D}) + c_2, \tag{16}$$

$$\lambda \;=\; 1-(1-f(\theta))(1-g(\bar{D})). \tag{17}$$

We tested several sets of function parameters satisfying the aforementioned four traces, and the one with the highest accuracy is $a_1 = 0.5$, $b_1 = -0.5$, $c_1 = 1 - 0.5\tanh(\pi/4)$, $a_2 = 0.7$, $b_2 = 5$, $c_2 = 0$. The corresponding plot is show in Fig. 6(d). Similar to *Case 3*, after the fusion scores s_i ($i = 1, \ldots, m$) of all the m subsets in a video clip are calculated, the average score (Equation (**??**)) is used for recognition.

For comparison, the 'optimal' λ values calculated from the test set $\mathbb{P}_1$ of Dataset A used in Section 4 are plotted in Fig. 6(e). The continuous range of $\bar{D}$ in $[0, 1]$ is divided into 10 bins with width 0.1. Thus the λ values compose a 5×10 matrix with each entry corresponding to a different pair of $(\theta, \bar{D})$. For each test sample in $\mathbb{P}_1$, 11 different values of λ, from 0 to 1 with the step 0.1, are respectively used to combine the gait score and face score. The λ value corresponding to the highest rank of the correct class label is recorded as the best λ corresponding to the $(\theta, \bar{D})$ pair of that test sample. Finally, the average value at each $(\theta, \bar{D})$ entry in the matrix are calculated as the 'optimal' λ. As can be seen that the surfaces in Fig. 6(b), (c) and (d) are increasingly similar to that of the 'optimal' λ in Fig. 6(e). More specifically, in Fig. 6(e), λ tends to decrease with the increase of θ and increase with the increase of $\bar{D}$, which is consistent with the principles used in *Case 2* to *Case 4*.

It can be seen in the progression from *Case 1* to *Case 4* that λ becomes more sensitive to the changes of θ and $\tilde{D}$. At the same time, more parameters need to be empirically determined. These parameters must be carefully tuned to suit a specific database. For a more general solution, machine learning techniques, such as neural networks, can be adopted to automatically learn the relationship between

the fusion scheme and the context. Consequently, such approaches can be easily applied to different databases.

3.3.2. Context-Aware Fusion Based on Neural Network

The basic assumption in Section 3.3.1 is the weighted sum fusion rule. The system responds to the changes of view angle and distance through adjusting the weight λ, and the way of adjusting λ is empirically determined. In this section, all of these assumptions are removed by using a neural network to learn how to fuse gait score and face score according to different view angles and subject-to-camera distance. This means both $\mathbf{K}$ and f_a are implied by the neural network.

First, as described in *Case 3* and *Case 4*, each video clip is also divided into several subsets and the average distance $\bar{D}$ is calculated. Then, for each subset, a feature vector $\boldsymbol{v}$ is composed

$$\boldsymbol{v} = \langle \tilde{s}_g^1, \tilde{s}_g^2, \ldots, \tilde{s}_g^N, \tilde{s}_f^1, \tilde{s}_f^2, \ldots, \tilde{s}_f^N, \bar{D}, \boldsymbol{u}(\theta) \rangle, \tag{18}$$

where $\tilde{s}_g^p$ and $\tilde{s}_f^p$ ($p = 1, \ldots, N$) are the normalized gait and face matching scores to person p respectively, and N is the number of known persons in the database. Since there are only 5 possible values of θ ($-\pi/2$, $-\pi/4$, 0, $\pi/4$, and $\pi/2$), a binary 5-dimensional vector $\boldsymbol{u}(\theta)$ is integrated into the feature vector. The i-th element in $\boldsymbol{u}(\theta)$ is 1 if θ equals the i-th value, and 0 otherwise. In the same way, the class label (personal ID) of each subset is also transformed into an N-dimensional binary vector $\boldsymbol{l}$. Thus for each subset in the training set, a feature-label pair $(\boldsymbol{v}, \boldsymbol{l})$ is generated and used to train a feedforward network with one hidden layer by the backpropagation algorithm [36]. The architecture of the network is shown is Fig. 7. There are $2N + 6$ (the dimensionality of $\boldsymbol{v}$) neurons in the input layer and N (the dimensionality of $\boldsymbol{l}$) neurons in the output layer. Details about how to de-

termine the number of neurons in the hidden layer will be described in Section 4.

After the neural network is trained, given a new video clip, it is first divided into m subsets as described before. For each subset, compose its feature vector $\boldsymbol{v}_i$ ($i = 1, \ldots, m$), input $\boldsymbol{v}_i$ into the neural network, get the output vector $\boldsymbol{l}_i$. The p-th element in $\boldsymbol{l}_i$, $\boldsymbol{l}_i^p$, can be viewed as a vote for person p. Then all $\boldsymbol{l}_i$'s are sum over all subsets and the largest element in the result indicates the ID of the person in the video clip:

$$ID = \arg\max_p (\sum_i \boldsymbol{l}_i^p). \tag{19}$$

In the whole procedure, there are no empirical assumptions on how the fusion should respond to the changes of context. All kinds of information including the gait scores, the face scores, the view angle, and the subject-to-camera distance is input into the neural network and we allow the learning algorithm to find out the relationship between them. Intuitively, this approach should perform better than those based on weighted sum because: (a) the knowledge $\mathbf{K}$ and the fusion function f_a are learned from a training set; (b) the relationship learned by the neural network is nonlinear, which is generally believed more suitable for realistic problems.

4. Experiment

4.1. Methodology

4.1.1. Data Sets

The data used in the experiment is the Dataset A (the former NLPR Gait Database [37]) and a subset of Dataset B in the CASIA Gait Database [21]. There are 20 different subjects in Dataset A. All the videos are captured in an outdoor environment. Each subject walks along a straight-line path back and forth twice

with three different angles between the path line and the image plane: lateral (0°), oblique (45°), and front/back (90°). Using the convention shown in Fig. 4, the view angles θ included in Dataset A are: $-\pi/2$, $-\pi/4$, 0, $\pi/2$, $3\pi/4$, and π. Since the left-right flipping hardly changes the relationship between gait and face in the fusion due to human body symmetry, $3\pi/4$ belongs to pattern (iv), and π belongs to pattern (iii). In total, there are 20 (persons) $\times$3 (path angles) $\times$2 (back and forth) $\times$2 (twice) = 240 gait video clips in Dataset A. Typical video frames from Dataset A in the five walking patterns (i-v) are shown in Fig. 8(a).

Dataset B is a much larger database, including 124 different subjects with variations in view angle and walking status (normal, in a coat, or with a bag). All the videos are captured in a well controlled laboratorial environment. Since clothing and carrying condition changes are not in the scope of this paper, only the videos with normal walking status are used in this experiment. Five typical view angles (000°, 036°, 090°, 144°, and 180°) corresponding to the five walking patterns shown in Fig. 4 are selected. Although the oblique views (036° and 144°) are not exactly $\pm\pi/4$, they can be accurately detected as pattern (iv) and (ii) respectively by the rules listed in Table 3. Thus the view angles can be roughly transformed into the convention shown in Fig. 4 as $\pi/2$, $3\pi/4$, π, $-3\pi/4$, and $-\pi/2$, respectively. By symmetry, the angle out of $[-\pi/2, \pi/2]$ can be mapped into one of the five patterns: $3\pi/4$ as pattern (iv), π as pattern (iii), and $-3\pi/4$ as pattern (ii). In this dataset, each subject walks along a straight line for 6 times while being captured by multiple cameras from all the view angles. In total, there are 124 (persons) $\times$5 (view angles) $\times$6 (times) = 3720 gait video clips in the experimental Dataset B. Typical video frames from Dataset B in the five walking patterns are shown in Fig. 8(b).

In both Dataset A and B, the videos of each individual in the same walking pattern are randomly divided into equal-sized training set (gallery set, denoted by $\mathbb{G}_1$) and test set (probe set, denoted by $\mathbb{P}_1$). Based on the training set, a gait score generator (Equation (4)) and a face score generator (Equation (7)) are trained. Then the recognition rates of the probe videos by the gallery video database based on gait-only, face-only, and the fusion of them are compared. The fusion methods include the context-aware fusion based on weighted sum or neural network proposed in Section 3, and those static fusion rules commonly used by most previous works on multi-biometric fusion [26] [27] [28], namely SUM, PRODUCT, MIN, and MAX.

4.1.2. Preprocessing and System Parameters

For gait score calculation described in Section 3.1.1, the silhouette images are normalized to 48×32, the subspace dimensionality $l = 25$ and the neighborhood size $k = 15$. For face score calculation described in Section 3.1.2, the face images are normalized to 25×25, the subspace dimensionality $q = N - 1$, where N is the number of different individuals in the database. When determining the walking patterns, the ratio threshold $\rho_1 = \rho_2 = 0.1$ in Table 3. In case the video clips need to be divided into subsets (*Case 3*, *Case 4* in the weight-sum-based fusion and the neural-network-based fusion), the number of frames in each subset $r = 30$ (since the longest walking cycle in the data sets is no more than 30 frames), and the overlap $v = 15$.

There are increasing number of parameters in *Case 2* to *Case 4* of the weighted-sum-based fusion, which need to be empirically determined. For each case, several representative parameter settings are tested and the best one is chosen. Fig. 9 illustrates how the parameters for *Case 2* is determined. Recall that in *Case 2*, λ is

determined only by the view angle (walking pattern). The representative parameter settings for *Case 2* must satisfy the following conditions: (a) λ decreases with the increase of view angle; (b) $\lambda = 1$ for walking pattern (i); (c) the parameter settings are significantly different from each other. In this case, condition (c) means λ decreases with significantly different rate. In Fig. 9, four settings are tested on Dataset A. The recognition accuracy of each setting is shown in the legend box, and the best one is chosen. The parameter settings for *Case 3* and *Case 4* are determined in the similar way, which have been given in Section 3.3.1.

For the context-aware fusion based on neural network, besides the training set for the gait score generator and the face score generator, one more training set is needed to train the neural network. Thus the test set needs to be further divided into two parts, one as training set (denoted by $\mathbb{T}_2$), the other as test set (denoted by $\mathbb{P}_2$). For Dataset A, because there is only one test video from each individual in each view angle, the test video is divided into subsets of length $r = 30$ frames, as described before. Then, half of these subsets are randomly selected, the corresponding feature vectors $\boldsymbol{v}$ (Equation (18)) and label vectors $\boldsymbol{l}$ are calculated, and the $(\boldsymbol{v}, \boldsymbol{l})$ pairs are used as the training set for the neural network shown in Fig. 7. The other half are used as the test set. Note that each video in this test set is only about half as long as the original video. For Dataset B, since there are three test videos from each individual in each view angle, two of them are randomly selected as training set for the neural network and the other is used as test set.

There is only one parameter in the neural-network-based fusion, i.e., the number of neurons κ in the hidden layer. In order to find a suitable value for κ, 10% of $\mathbb{G}_2$ are randomly selected as a validation set to test the neural network trained on the remaining 90% of $\mathbb{G}_2$. Ten-fold cross validations are performed for different

values of κ. Finally, the κ value with the highest average accuracy is selected as the best setting. In this experiment, the best setting of κ for Dataset A is 160, and that for Dataset B is 400. Note that in the whole procedure of seeking the best κ, the test set $\mathbb{P}_2$ is not used, which ensures a fair comparison on $\mathbb{P}_2$ with other algorithms.

4.2. Results

4.2.1. Single Modality Versus Multi-modality Fusion

The recognition rates of gait-only, face-only, the static fusion rules (SUM, PRODUCT, MIN, and MAX), the four cases of weighted-sum-based context-aware fusion, as well as the weighted sum fusion using the 'optimal' λ plotted in Fig. 6(e) on Dataset A (trained on $\mathbb{G}_1$ and tested on $\mathbb{P}_1$) are tabulated in Table 4 and those on Dataset B are tabulated in Table 5 The recognition rates are calculated for all the test videos in $\mathbb{P}_1$, as well as each walking pattern separately. The best performance in each case is highlighted by boldface, and the fusion results better than both gait-only and face-only are underlined. The results of gait-only and face-only in different walking patterns support the prior knowledge about the relationship between gait/face recognition and the view angle used in Section 3.3.1, i.e., gait recognition prefers the side view (walking pattern (iii)) while face recognition performs better in the frontal view (walking pattern (v)). It might be a little strange to see a 55% face recognition rate for the back view (walking pattern (i)) when there is no face actually shown in the images. The possible reason is that the hair can also be used to distinguish different people [38]. It can be seen that none of the static fusion rules can guarantee a better result than the single biometric traits in all cases. As for the overall performance, only SUM and MAX can get slightly better results than both gait-only and face-only. This is due to the

Table 4: Recognition Rates (%) on the Test Set $\mathbb{P}_1$ of Dataset A

Walking Pattern	(i)	(ii)	(iii)	(iv)	(v)	All
Gait-only	50.0	50.0	72.5	65.0	60.0	61.7
Face-only	55.0	55.0	57.5	65.0	70.0	60.0
SUM	75.0	55.0	80.0	60.0	60.0	68.3
PRODUCT	75.0	45.0	72.5	35.0	45.0	57.5
MIN	80.0	30.0	65.0	30.0	50.0	53.3
MAX	70.0	75.0	57.5	70.0	60.0	65.0
Case 1	80.0	95.0	70.0	**95.0**	85.0	82.5
Case 2	80.0	**100.0**	85.0	**95.0**	80.0	87.5
Case 3	80.0	**100.0**	**85.5**	85.0	85.0	88.3
Case 4	80.0	**100.0**	82.5	**95.0**	90.0	88.3
'Optimal' λ	**85.0**	**100.0**	82.5	**95.0**	**95.0**	**90.0**

usage of the fixed fusion rules without considering the reliability of different biometric traits under different conditions. An unreliable single biometric trait might deteriorate the performance of the other better one in the fusion.

With the ability to perceive view angle (walking pattern) and subject-to-camera distance, all the four cases of the weighted-sum-based context-aware fusion perform not only significantly better than both single biometric traits, but also significantly better than all the static fusion rules. Specially, *Case 1* also uses the static SUM rule. Thus its superiority over directly applying the SUM rule mainly comes from the way of calculating the matching scores according to different walking patterns. This also suggests possibilities for further improvement by applying the

Table 5: Recognition Rates (%) on the Test Set $\mathbb{P}_1$ of Dataset B

Walking Pattern	(i)	(ii)	(iii)	(iv)	(v)	All
Gait-only	65.0	68.3	71.4	70.3	69.2	68.8
Face-only	63.3	65.3	66.7	66.4	71.4	66.6
SUM	66.4	68.6	70.6	68.6	71.9	69.2
PRODUCT	58.9	58.3	69.4	67.8	63.3	63.6
MIN	67.5	53.9	67.8	64.7	69.4	64.7
MAX	66.1	70.3	75.0	68.9	71.4	70.3
Case 1	83.3	89.4	61.4	72.2	77.5	76.8
Case 2	71.7	87.2	88.6	89.4	83.1	84.0
Case 3	72.2	88.3	89.2	91.1	86.7	85.5
Case 4	85.6	90.0	89.4	91.1	86.7	88.6
'Optimal' λ	**86.9**	**91.11**	**91.7**	**92.5**	**86.9**	**89.8**

techniques for view-based gait recognition [39, 40, 41] and/or view-based face recognition [42]. From *Case 2* to *Case 4*, the fusion is more and more sensitive to the changes of context. Thus better performance than *Case 1* can be achieved. Since both *Case 3* and *Case 4* can dynamically adjust the fusion weights according to both view angle and subject-to-camera distance, they achieve the highest overall accuracy among the four context-aware fusion cases, which is only slightly lower than that of the fusion using 'optimal' λ. Note that the 'optimal' λ is calculated from the test set $\mathbb{P}_1$. Thus the corresponding accuracy can be viewed as the upper bound of the accuracy of the fusion schemes based on weighted sum tested on $\mathbb{P}_1$. It is also notable that *Case 4* performs better than the 'optimal' λ for the

walking pattern (iii). This is because that the 'optimal' λ is not the real optimal λ since: 1. the 'optimal' λ is obtained through roughly quantizing the continuous distance $\bar{D}$ and λ with the width 0.1; 2. the real optimal λ for each test sample should be different, but the 'optimal' λ is an average value for all the samples in the test set $\mathbb{P}_1$. Thus, although the 'optimal' λ generally should achieve the best performance, it is not guaranteed.

Note that Dataset B is much larger than Dataset A, but the overall performance of the context-aware fusion schemes on Dataset B is not significantly worse than that on Dataset A, some are even better, such as *Case 4*. This shows the good generalization ability of context-aware fusion. Another reason might be that Dataset B is obtained in a well controlled laboratorial environment (e.g., relatively steady illumination, predefined walking route, precise view angle, etc.), thus different videos of the same person in the same view angle might be more similar than those videos in Dataset A obtained in outdoor environment. This also explains why Gait-only and Face-only perform better on Dataset B than on Dataset A.

4.2.2. Prior-knowledge-based Fusion Versus Machine-learning-based Fusion

The context-aware fusion schemes based on weighted sum (including the 'optimal' λ) and the neural-network-based fusion are further tested on $\mathbb{P}_2$ of Dataset A and B, and the results are compared in Table 6 and Table 7, respectively. The highest recognition rate in each case is highlighted by boldface. Note that it is not appropriate to test the neural network on $\mathbb{P}_1$ since half of the video subsets in it ($\mathbb{T}_2$) were used for the training of the neural network. As can be seen, from *Case 1* to *Case 4*, the overall performance becomes increasingly better since they are more and more sensitive to view angle and subject-to-camera distance. The weighted sum fusion using the 'optimal' λ still gets slightly higher accuracy than

Table 6: Recognition Rates (%) on the Test Set $\mathbb{P}_2$ of Dataset A

Walking Pattern	(i)	(ii)	(iii)	(iv)	(v)	All
Case 1	80.0	80.0	75.7	85.0	80.0	79.5
Case 2	80.0	90.0	86.5	95.0	80.0	86.3
Case 3	80.0	85.0	86.5	95.0	90.0	87.2
Case 4	85.0	95.0	89.2	95.0	80.0	88.9
'Optimal' λ	85.0	95.0	89.2	95.0	85.0	89.7
Neural Network	**100.0**	**100.0**	**97.3**	**100.0**	**100.0**	**99.2**

the four cases. The neural-network-based fusion achieves very high accuracy on $\mathbb{P}_2$. The significant improvement over the weighed-sum-based fusion is due to the ability of the neural network to learn the knowledge $\mathbf{K}$ and the fusion function f_a from the training set $\mathbb{T}_2$, rather than empirically determining them like that in *Case 1* to *Case 4*. Specially, the accuracy of the neural-network-based fusion is much higher than the upper bound of the weighed-sum-based fusion represented by the 'optimal' λ. This indicates the advantages of the neural network to model the complex relationship among the individual biometrics and the context factors, rather than assume the relationship to be as simple as a weighted sum. However, the disadvantage of the neural-network-based fusion is in its scalability: when the persons in the database change, the neural network must be re-trained. The training process might be time-consuming when the database is very large. Thus for applications with a frequently variational large database, the fusion based on weighted sum (*Case 1* to *Case 4*) might be a more practical choice.

Table 7: Recognition Rates (%) on the Test Set $\mathbb{P}_2$ of Dataset B

Walking Pattern	(i)	(ii)	(iii)	(iv)	(v)	All
Case 1	80.0	80.8	72.5	75.0	84.2	78.5
Case 2	69.2	87.5	82.5	75.0	84.2	79.7
Case 3	69.2	87.5	82.5	85.0	90.8	83.0
Case 4	85.8	87.5	83.3	80.8	84.2	84.3
'Optimal' λ	80.8	89.2	90.0	91.7	85.0	87.3
Neural Network	**99.2**	**97.5**	**95.8**	**100.0**	**100.0**	**98.5**

4.2.3. Effects of External Factors

In order to reveal the influence of the subject-to-camera distance, each subset of the videos in $\mathbb{P}_1$ of Dataset A is independently recognized, rather than combining the results of all subsets in one video (Equation (**??**)). These subsets are then grouped by their corresponding average distance $\bar{D}$. The possible range of $\bar{D}$, $[0, 1]$, is equally divided into 5 bins, i.e., $[0, 0.2)$, $[0.2, 0.4)$, $[0.4, 0.6)$, $[0.6, 0.8)$, and $[0.8, 1]$. The subsets are selected into the same group if their corresponding $\bar{D}$'s fall into the same bin. Thus the 5 groups represent the video subsets with different subject-to-camera distance. The recognition rate in each group is calculated for gait-only, face-only and the four cases of the weighted-sum-based context-aware fusion. The results are shown in Fig. 10. As can be seen that the further the subject to the camera, the worse face-only performs. The recognition rate of gait-only does not always decrease with the increase of $\bar{D}$, which indicates that it is not apparently related to $\bar{D}$. But when the subject is far away ($\bar{D} \in [0.8, 1]$), the motion characters might be difficult to capture, thus gait-only also performs

poorly. Combined with the worst performance of face-only at $[0.8, 1]$, it is not surprising to see the accuracy of all the fusion methods deteriorates dramatically at this range. All the four cases of the context-aware fusion based on weighted sum perform better than both single biometric traits at all distance ranges, among which *Case 3* and *Case 4* achieve the best results because they can adapt the fusion weights to both view angle and subject-to-camera distance.

In order to test the sensitivity of the neural-network-based context-aware fusion to the limitations of the measurement process, only the video segments with certain range of subject-to-camera distance are used to train the neural network, but the trained neural network is tested by video segments with full range of distance. The result on Dataset B is shown in Fig. 11. The neural network is trained by the video segments in $\mathbb{T}_2$ with $\bar{D}$ in $[0, 1]$ (full range), $[0, 0.8]$, $[0, 0.6]$, $[0, 0.4]$, and $[0, 0.2]$, respectively, and then tested by all the videos in the the test set $\mathbb{P}_2$. As can be seen, the performance of the neural network remains almost the same when the training distance range is reduced to $[0, 0.8]$, but starts to deteriorate significantly after the training distance range shrinks to $[0, 0.6]$ or smaller. Since the measurement process is not likely to be too restricted (such as only able to measure the distance within $[0, 0.8]$), the neural-network-based fusion can be regarded as not sensitive to the measurement limitations.

In order to test the sensitivity of the neural-network-based fusion to novel context, only part of the training set $\mathbb{T}_2$ of Dataset B are used to train the neural network (thus the context variations included in the training set are reduced), and then it is tested on all the videos in $\mathbb{P}_2$ of Dataset B (thus the test set still includes all context variations). The percentage of training samples out of $\mathbb{G}_2$ decreases from 100% to 20%, with the step 20%. The result is shown in Fig. 12. As can

be seen that the recognition rate of the neural network remains higher than 90% even when only 40% samples of the training set are used. The insensitivity of the proposed fusion scheme to training samples illustrates its ability to deal with novel context.

In order to test the scalability of the proposed context-aware fusion schemes, Dataset B is divided into four subsets. The first is a *generic training set* (denoted by $\mathbb{T}$) including all the videos of half subjects in Dataset B (62 persons). The other three subsets are from the videos of the remaining 62 subjects in Dataset B, where the six videos of each individual in the same walking pattern are equally divided into a *gallery set* (denoted by $\mathbb{G}_3$) and a *probe set*. In order to train the neural network, the three videos of each individual in the same walking pattern in the probe set are further divided into a training set (denoted by $\mathbb{T}_3$) containing two videos and a test set (denoted by $\mathbb{P}_3$) containing the remaining one. $\mathbb{T}$ is used to train the gait and face classifiers described in Section 3.1. $\mathbb{G}_3$ is used to extract the personal features by the classifiers. $\mathbb{P}_3$ is used to test the fusion schemes, i.e., find the most similar person in $\mathbb{G}_3$ for each video in $\mathbb{P}_3$. $\mathbb{T}_3$ is used only for the neural-network-based fusion, but not for the weighted-sum-based fusion. In this way, the weighted-sum-based fusion schemes are trained on $\mathbb{T}$ and tested on previously unseen subjects. As for the neural-network-based fusion, as mentioned, it requires one more training on $\mathbb{T}_3$ from the new subjects, but the gait and face classifier are still trained on different subjects ($\mathbb{T}$). The results are tabulated in Table 8. Compared with Table 7, Table 8 shows even better performance. Besides the reason that the number of subjects to recognize is reduced to half, this also indicates good scalability of the proposed methods. When applied to previously unseen subjects, the weighted-sum-based schemes can be directly used without

Table 8: Recognition Rates (%) on the Test Set $\mathbb{P}_3$ of Dataset B

Walking Pattern	(i)	(ii)	(iii)	(iv)	(v)	All
Case 1	86.7	81.7	71.7	85.0	88.3	82.7
Case 2	75.0	85.0	83.3	86.7	88.3	83.7
Case 3	75.0	86.7	83.3	88.3	91.7	85.0
Case 4	86.7	86.7	83.3	90.0	88.3	87.0
'Optimal' λ	88.3	90.0	91.7	95.0	88.3	90.7
Neural Network	**98.3**	**96.7**	**98.3**	**100.0**	**100.0**	**98.7**

re-training. For the neural-network-based scheme, only the neural network itself need to be trained to fit the new data, while the existing gait and face score generators do not need to be re-trained.

5. Conclusion and Discussion

This paper proposes *context-aware multi-biometric fusion*. Up to the present, most existing work on multi-biometric fusion is based on *static fusion*. On the contrary, context-aware fusion can perceive the changes of the external factors and dynamically adapt the fusion rule to those changes. To illustrate the advantages of context-aware fusion, the fusion of gait and face in video which is adaptable to view angle and subject-to-camera distance is investigated. Several context-aware fusion schemes based on either prior knowledge (weighted sum) or machine learning (neural network) are proposed. Experimental results reveal that the context-aware fusion methods perform significantly better than conventional static fusion rules, such as SUM, PRODUCT, MIN and MAX. Moreover,

the context-aware fusion based on machine learning can remarkably improve that based on prior knowledge.

The flowchart of context-aware multi-biometric fusion shown in Fig. 1 is a general framework, which leaves a lot of possibilities for the future work that might further improve the results reported in this paper. First, for the fusion level, although only the score level fusion is investigated in this paper, other fusion levels, such as the more essential feature level fusion could also be adopted in the framework. As for the fusion rules, other than the empirically determined weighted sum fusion and the neural network based fusion, other optimization methods, such as the Genetic Algorithm and SVM, could be applied to find out the relationship between the context factors and the fusion rules.

Second, the detection methods for the context factors could be further improved. While this paper assumes people walking along a straight path (which is the most common case) with five quantized angles, in practice, people might change their paths unpredictably. Thus in order to make the system effective in more realistic scenarios, more sophisticated techniques could be adopted for pose estimation, such as tracking the location of the head [43, 44], or face pose estimation [45, 46, 47]. Moreover, additional special sensors can also help to improve the performance, such as using a laser distance sensor to detect the subject-to-camera distance.

Third, this paper studies one typical application of context-aware multi-biometric fusion, i.e., the adaptive fusion of gait and face in video for human identification. In the future, the context-aware fusion of other biometric traits (fusion of two or more biometrics) will be further investigated, such as the adaptive fusion of face, fingerprint, iris, and voice, etc. Also, the application of the context-aware fusion

to a verification scenario is also possible with a few changes in the design of the fusion rules, e.g., change the output layer of the neural network shown in Fig. 7 into just two neurons, one for impostor scores, and the other for genuine scores. For realistic surveillance systems, open-set identification should be implemented, i.e., add one more class as 'not in the list'. This could be realized by adding a final verification step. One simple implementation is to classify the fused score lower than a certain threshold as 'not in the list'. More sophisticated methods for the open-set problem in context-aware fusion need further investigation.

With the development of sensor techniques, the collection of multiple biometric traits from one subject is becoming easier. When there are a number of biometric traits available, the problem of selecting an optimal set of traits for fusion rises. The selection could also be adaptive to external conditions. For example, under poor illumination condition, the visual biometric traits (e.g., face) should not be selected, and in noisy environment, the sound-based traits (e.g., voice) should not be selected. Moreover, the diversity among the available biometric traits is also an important issue to be considered when making selections. Thus the framework of context-aware multi-biometric fusion could be extended by adding a *Context-Aware Biometric Selection* module. This is also an important part of the further work.

Acknowledgment

This work was partially supported by the Australian Research Council Discovery Grant (DP0987421), the National Science Foundation of China (60905031), and the Jiangsu Science Foundation (BK2009269). The authors would like to thank the associate editor, Professor John Illingworth, and the anonymous review-

ers for their comments and suggestions which greatly improved this paper.

References

[1] A. K. Jain, A. Ross, Multibiometric systems, Communications of the ACM, Special Issue on Multimodal Interfaces 47 (1) (2004) 34–40.

[2] L. Hong, A. K. Jain, S. Pankanti, Can multibiometrics improve performance?, in: Proc. IEEE Workshop on Automatic Identification Advanced Technologies, NJ, USA, 1999, pp. 59–64.

[3] K. I. Chang, K. W. Bowyer, P. J. Flynn, An evaluation of multimodal 2d+3d face biometrics, IEEE Trans. Pattern Anal. Mach. Intell. 27 (4) (2005) 619–624.

[4] A. Ross, A. K. Jain, J. Reisman, A hybrid fingerprint matcher, Pattern Recognition 36 (7) (2003) 1661–1673.

[5] A. Ross, A. K. Jain, Multimodal biometrics: An overview, in: Proc. European Signal Processing Conference, Vienna, Austria, 2004, pp. 1221–1224.

[6] R. Brunelli, D. Falavigna, Person identification using multiple cues, IEEE Trans. Pattern Anal. Mach. Intell. 17 (10) (1995) 955–966.

[7] L. Hong, A. K. Jain, Integrating faces and fingerprints for personal identification, IEEE Trans. Pattern Anal. Mach. Intell. 20 (12) (1998) 1295–1307.

[8] A. K. Jain, K. Nandakumar, A. Ross, Score normalization in multimodal biometric systems, Pattern Recognition 38 (12) (2005) 2270–2285.

[9] A. K. Dey, Understanding and using context, Personal and Ubiquitous Computing 5 (1) (2001) 4–7.

[10] A. Schmidt, M. Beigl, H.-W. Gellersen, There is more to context than location, Computers & Graphics 23 (6) (1999) 893–901.

[11] S. Chu, M. Yeung, L. Liang, X. Liu, Environment-adaptive multi-channel biometrics, in: Proc. Int'l Conf. Acoustics, Speech and Signal Processing, Hong Kong, China, 2003, pp. V–788–791.

[12] J. Fiérrez-Aguilar, Y. Chen, J. Ortega-Garcia, A. K. Jain, Incorporating image quality in multi-algorithm fingerprint verification, in: Proc. Int'l Conf. Biometrics, Hong Kong, China, 2006, pp. 213–220.

[13] J. Fiérrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, J. Bigün, Discriminative multimodal biometric authentication based on quality measures, Pattern Recognition 38 (5) (2005) 777–779.

[14] H. P.-S. Hui, H. M. Meng, M.-W. Mak, Adaptive weight estimation in multi-biometric verification using fuzzy decision fusion, in: Proc. Int'l Conf. Acoustics, Speech and Signal Processing, Honolulu, Hawaii, 2007, pp. I–501–504.

[15] O. Fatukasi, J. Kittler, N. Poh, Quality controlled multimodal fusion of biometric experts, in: Proc. 12th Iberoamerican Congress on Pattern Recognition, LNCS 4756, Valparaiso, Chile, 2007, p. 881C890.

[16] K. Nandakumar, Y. Chen, A. K. Jain, S. C. Dass, Quality-based score level fusion in multibiometric systems, in: Proc. Int'l Conf. Pattern Recognition, Hong Kong, China, 2006, pp. 473–476.

[17] U. Park, A. K. Jain, A. Ross, Face recognition in video: Adaptive fusion of multiple matchers, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, Minneapolis, MN, 2007, pp. 1–8.

[18] N. Poh, S. Bengio, Improving fusion with margin-derived confidence in biometric authentication tasks, in: Proc. Int'l Conf. Audio- and Video-Based Biometric Person Authentication, Hilton Rye Town, NY, 2005, pp. 474–483.

[19] INCITS Project 1672-D, Biometric sample quality standard draft (revision 4), Tech. Rep. INCITS/M1/06-0948, InterNational Committee for Information Technology Standards, Washington, DC (November 2006).

[20] A. Krogh, J. Vedelsby, Neural network ensembles, cross validation, and active learning, in: G. Tesauro, D. S. Touretzky, T. K. Leen (Eds.), Advances in Neural Information Processing Systems 7, Denver, CO, 1995, pp. 231–238.

[21] CASIA Gait Database: http://www.cbsr.ia.ac.cn/Gait%20Database.htm.

[22] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, K. W. Bowyer, The humanid gait challenge problem: Data sets, performance, and analysis, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2) (2005) 162–177.

[23] The MIT gait recognition database. http://www.ai.mit.edu/people/llee/HID /data.htm.

[24] R. Gross, J. Shi, The CMU motion of body (mobo) database, Tech. Rep. CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA (June 2001).

[25] L. Xu, A. Krzyzak, C. Suen, Methods of combining multiple classifiers and their applications tohandwriting recognition, IEEE Transactions on Systems, Man, and Cybernetics 22 (3) (1992) 418–435.

[26] G. Shakhnarovich, T. Darrell, On probabilistic combination of face and gait cues for identification, in: Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition, Washington, DC, 2002, pp. 169–174.

[27] A. Kale, A. Roychowdhury, R. Chellappa, Fusion of gait and face for human identification, in: Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing, Montreal, Canada, 2004, pp. V–901–904.

[28] X. Zhou, B. Bhanu, Integrating face and gait for human recognition, in: Proc. Conf. Computer Vision and Pattern Recognition Workshop, New York City, NY, 2006, p. 55.

[29] X. Zhou, B. Bhanu, Integrating face and gait for human recognition at a distance in video, IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics 37 (5) (2007) 1119–1137.

[30] Z. Liu, S. Sarkar, Outdoor recognition at a distance by fusing gait and face, Image Vision Comput. 25 (6) (2007) 817–832.

[31] L. Wang, D. Suter, Analysing human movements from silhouettes using manifold learning, in: Proc. IEEE Int'l Conf. Advanced Video- and Signal-based Surveillance, Sydney, Australia, 2006, p. 7.

[32] M.-H. Yang, D. J. Kriegman, N. Ahuja, Detecting faces in images: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 24 (1) (2002) 34–58.

[33] D. Winter, Biomechanics and Motor Control of Human Movement, 2nd Ed., John Wiley & Sons, New York, NY, 1990.

[34] X. He, P. Niyogi, Locality preserving projections, in: Advances in Neural Information Processing Systems 16, British Columbia, Canada, 2003.

[35] P. N. Belhumeur, J. P. Hespanha, D. J. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, IEEE Trans. Pattern Anal. Machine Intell. 19 (7) (1997) 711–720.

[36] P. Werbos, Beyond regression: New tools for prediction and analysis in the behavioral sciences, Ph.D. thesis, Harvard University, Cambridge, MA (1974).

[37] L. Wang, T. Tan, H. Ning, W. Hu, Silhouette analysis-based gait recognition for human identification, IEEE Trans. Pattern Anal. Machine Intell. 25 (12) (2003) 1505– 1518.

[38] Y. Yacoob, L. S. Davis, Detection and analysis of hair, IEEE Trans. Pattern Anal. Mach. Intell. 28 (7) (2006) 1164–1169.

[39] N. Spencer, J. Carter, Towards pose invariant gait reconstruction, in: Proc. Int'l Conf. Image Processing, Genoa, Italy, 2005, pp. 261–264.

[40] M. Goffredo, R. D. Seely, J. N. Carter, M. S. Nixon, Markerless view independent gait analysis with self-camera calibration, in: Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition, Amsterdam, The Netherlands, 2008, pp. 1–6.

[41] S. Yu, D. Tan, T. Tan, Modelling the effect of view variation on appearance-based gait recognition, in: Proc. Asian Conf. Computer Vision, Hyderabad, India, 2006, pp. 807–816.

[42] Y. Li, S. Gong, J. Sherrah, H. M. Liddell, Support vector machine based multi-view face detection and recognition, Image Vision Comput. 22 (5) (2004) 413–427.

[43] L. M. G. Brown, 3d head tracking using motion adaptive texture-mapping, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, Kauai, HI, 2001, pp. 998–1003.

[44] W. Ryu, D. Kim, Robust 3d head tracking and its applications, in: Proc. Int'l Conf. Biometrics, Amsterdam, The Netherlands, 2007, pp. 968–977.

[45] Q. Ji, 3d face pose estimation and tracking from a monocular camera, Image Vision Comput. 20 (7) (2002) 499–511.

[46] V. Lepetit, J. Pilet, P. Fua, Point matching as a classification problem for fast and robust object pose estimation, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, Washington, DC, 2004, pp. 244–250.

[47] S. Srinivasan, K. L. Boyer, Head pose estimation using view based eigenspaces, in: Proc. Int'l Conf. Pattern Recognition, Quebec City, Canada, 2002, pp. 302–305.

Figure Captions

Fig. 1. Framework of context-aware multi-biometric fusion.

Fig. 2. The five representative walking patterns in the Dataset A of the CASIA Gait Database with indicative fusion weights.

Fig. 3. Gait and face extraction: (a) the original video image, (b) the background image, (c) the extracted silhouette, (d) extraction of the face image.

Fig. 4. View angle θ corresponding to each walking pattern.

Fig. 5. Estimation of the subject-to-camera distance.

Fig. 6. Variation of λ with θ and $\tilde{D}$: (a) *Case 1*, λ is constant, (b) *Case 2*, λ is constant for each view angle, (c) *Case 3*, λ varies with $\tilde{D}$ within a certain range determined by θ, (d) *Case 4*, λ is generated by a function of θ and $\tilde{D}$, and (e) the 'optimal' λ on $\mathbb{P}_1$ of Dataset A.

Fig. 7. Architecture of the neural network for context-aware fusion of gait and face.

Fig. 8. Typical video frames in the five walking patterns from (a) Dataset A and (b) Dataset B.

Fig. 9. Experiments on representative parameter settings for *Case 2* on Dataset A.

Fig. 10. Recognition rates on the subsets with different subject-to-camera distance.

Fig. 11. Recognition rate of the neural-network-based fusion with different range of $\bar{D}$ in the training set.

Fig. 12. Recognition rate of the neural-network-based fusion with different percentage of $\mathbb{T}_2$ as training set.

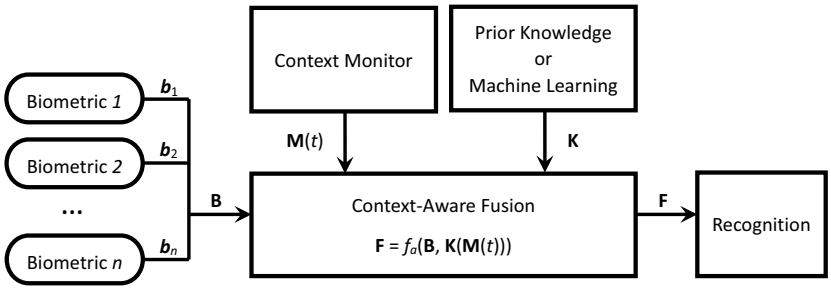


Figure 1: Framework of context-aware multi-biometric fusion.

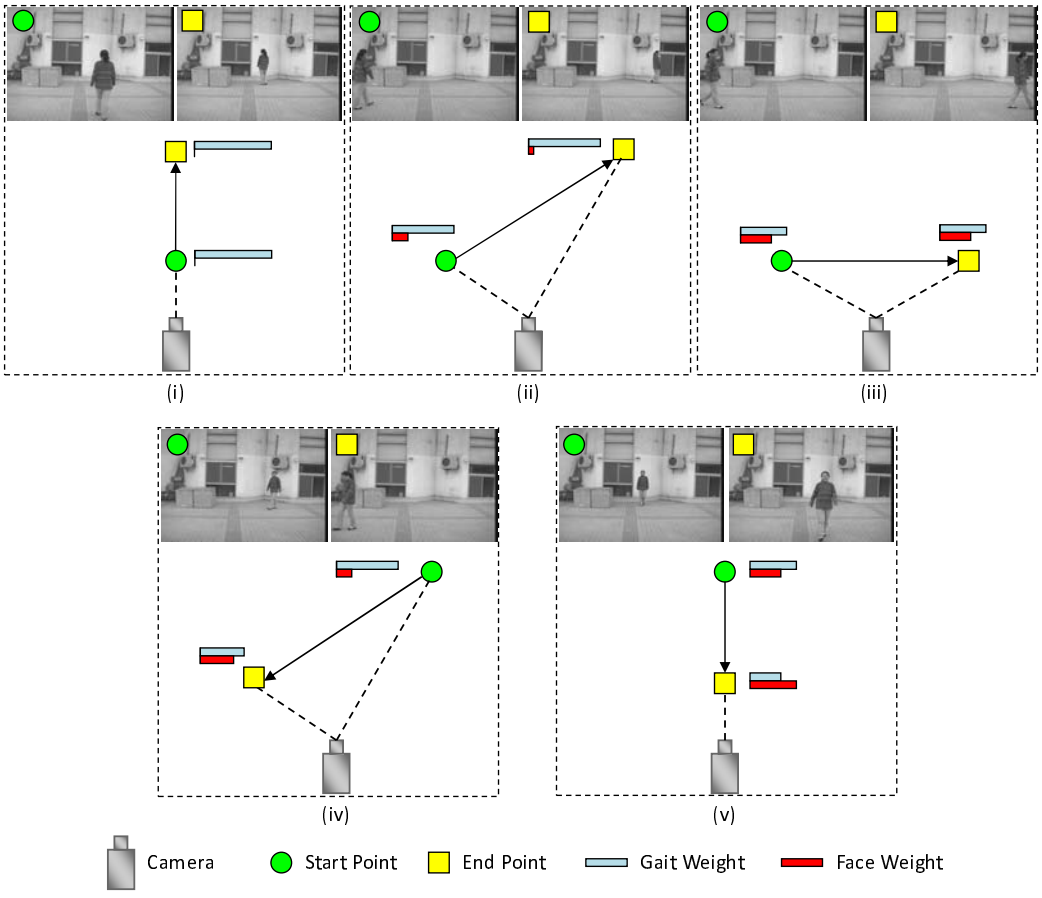
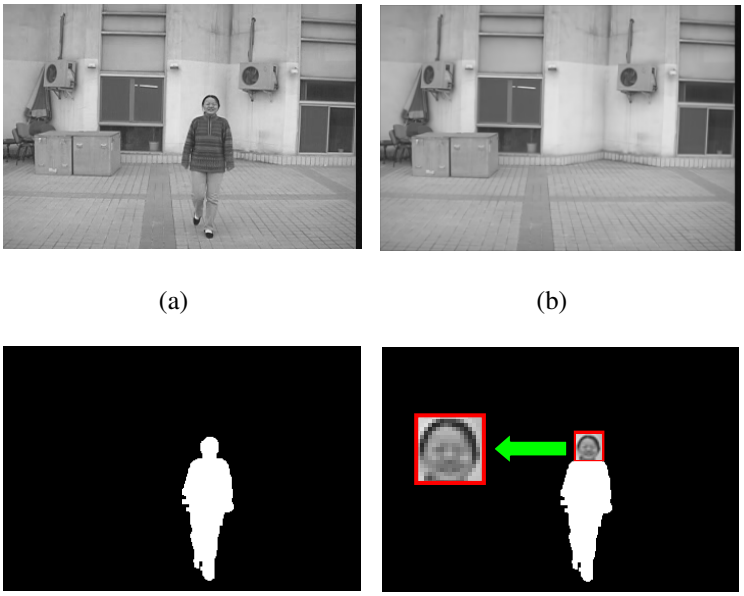


Figure 2: The five representative walking patterns in the Dataset A of the CASIA Gait Database with indicative fusion weights.



(a) (b)

(c) (d)

Figure 3: Gait and face extraction: (a) the original video image, (b) the background image, (c) the extracted silhouette, (d) extraction of the face image.

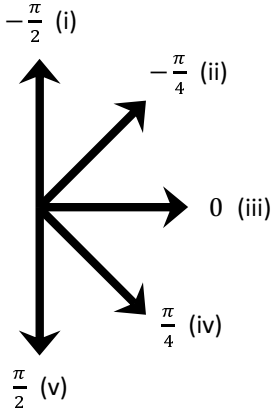


Figure 4: View angle θ corresponding to each walking pattern.

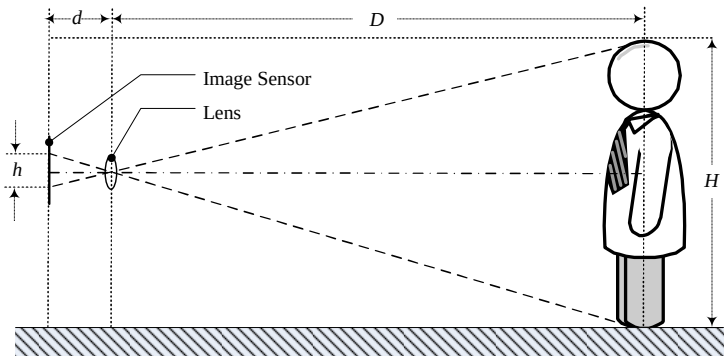


Figure 5: Estimation of the subject-to-camera distance.

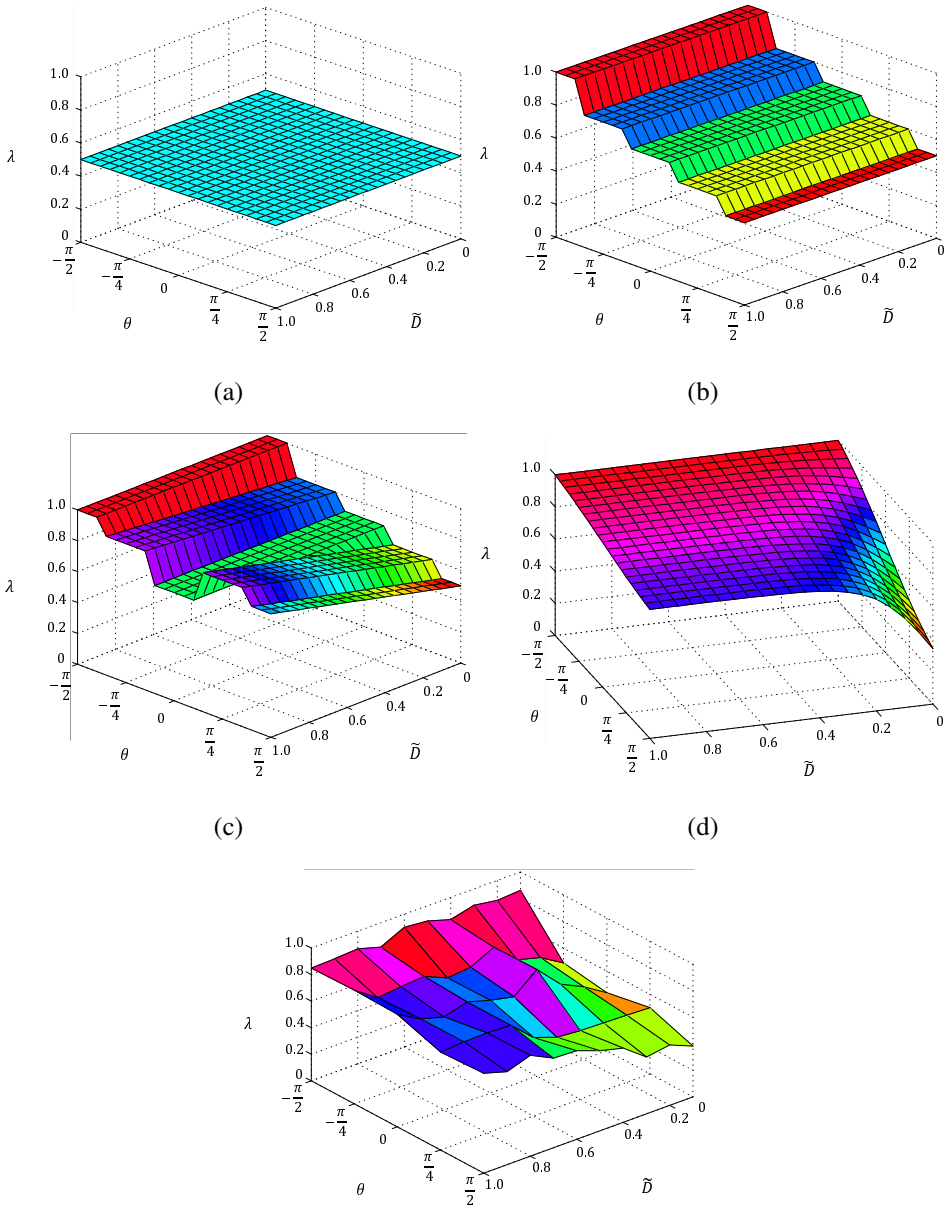


Figure 6: Variation of λ with θ and $\tilde{D}$: (a) *Case 1*, λ is constant, (b) *Case 2*, λ is constant for each view angle, (c) *Case 3*, λ varies with $\tilde{D}$ within a certain range determined by θ, (d) *Case 4*, λ is generated by a function of θ and $\tilde{D}$, and (e) the 'optimal' λ on $\mathbb{P}_1$ of Dataset A.

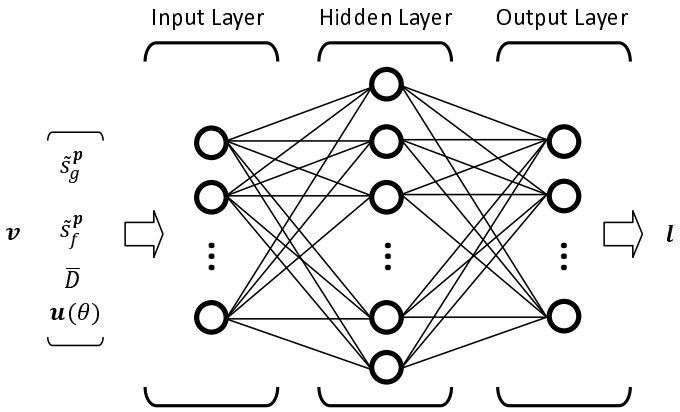
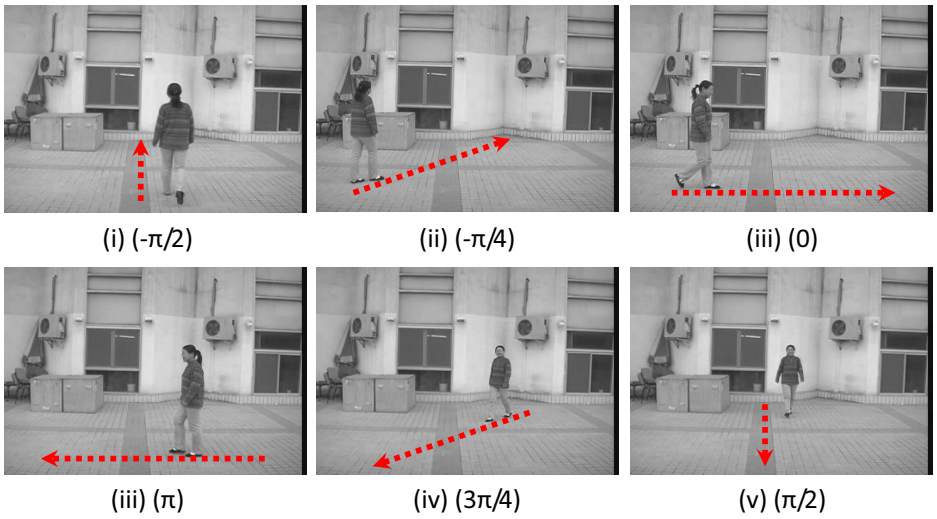
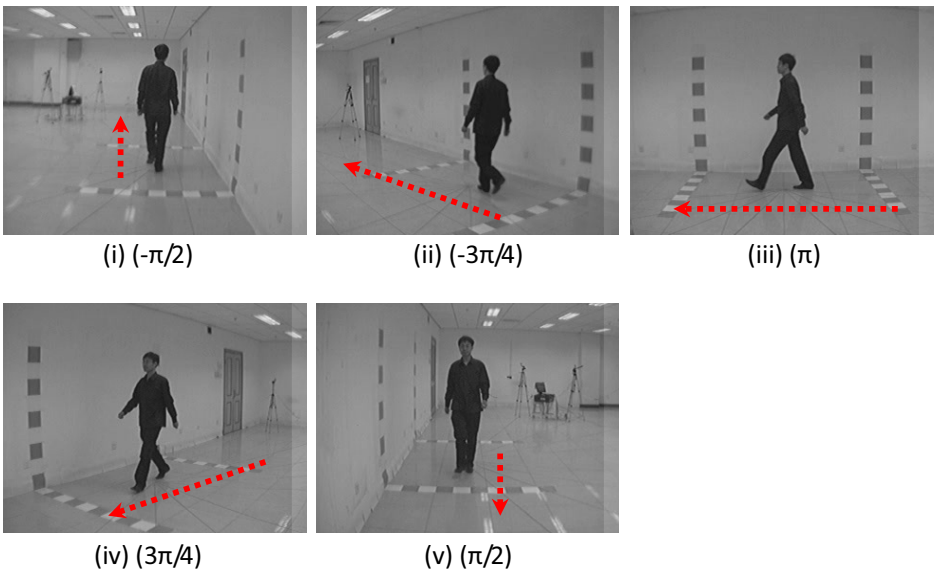


Figure 7: Architecture of the neural network for context-aware fusion of gait and face.



(a)



(b)

Figure 8: Typical video frames in the five walking patterns from (a) Dataset A and (b) Dataset B.

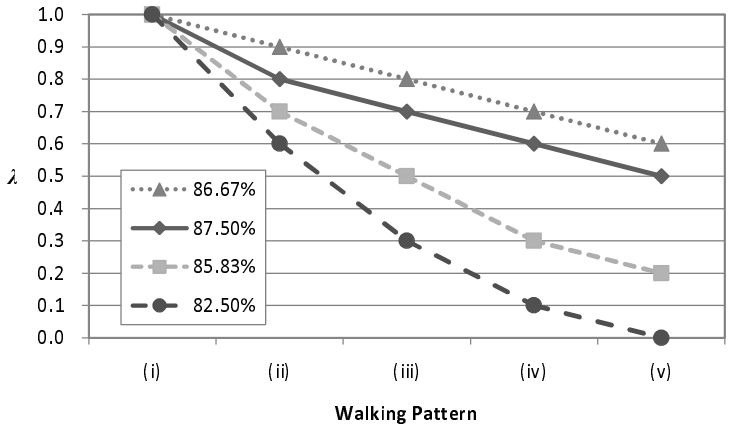


Figure 9: Experiments on representative parameter settings for *Case 2* on Dataset A.

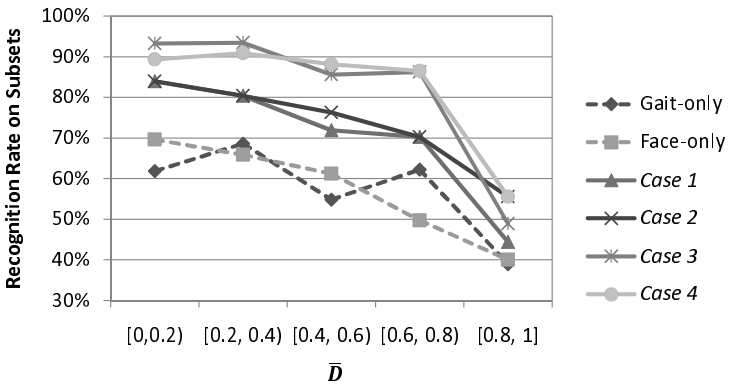


Figure 10: Recognition rates on the subsets with different subject-to-camera distance.

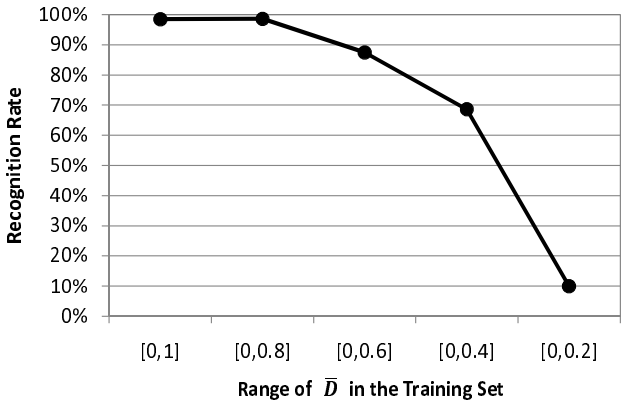


Figure 11: Recognition rate of the neural-network-based fusion with different range of $\bar{D}$ in the training set.

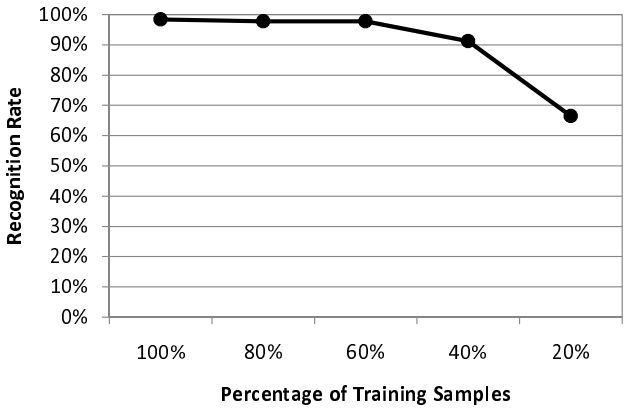


Figure 12: Recognition rate of the neural-network-based fusion with different percentage of $\mathbb{T}_2$ as training set.