Chapter XIX: Multi-Modal Biometrics Fusion for Human Recognition in Video

Xiaoli Zhou, University of California - Riverside, USA
Bir Bhanu, University of California - Riverside, USA

This chapter introduces a new video based recognition system to recognize non-cooperating individuals at a distance in video, who expose side views to the camera. Information from two biometric sources, side face and gait, is utilized and integrated for recognition. For side face, an Enhanced Side Face Image (ESFI), a higher resolution image compared with the image directly obtained from a single video frame, is constructed, which integrates face information from multiple video frames. For gait, the Gait Energy Image (GEI), a spatio-temporal compact representation of gait in video, is used to characterize human walking properties. The features of face and gait are extracted from ESFI and GEI, respectively. They are integrated at both of the match score level and the feature level by using different fusion strategies. The system is tested on a database of video sequences, corresponding to 45 people, which are collected over several months. The performance of different fusion methods are compared and analyzed. The experimental results show that (a) the idea of constructing ESFI from multiple frames is promising for human recognition in video and better face features are extracted from ESFI compared to those from the original side face images; (b) the synchronization of face and gait is not necessary for face template ESFI and gait template GEI. (c) integrated information from side face and gait is effective for human recognition in video. The feature level fusion methods achieve better performance than the match score level methods fusion overall.

I. Introduction

It has been found to be difficult to recognize a person from arbitrary views when one is walking at a distance. For optimal performance, a system should use as much information as possible from the observations. A fusion system, which combines face and gait cues from video sequences, is a potential approach to accomplish the task of human recognition. The general solution to analyze face and gait video data from arbitrary views is to estimate 3-D models. However, the problem of building reliable 3-D models for non-rigid face with flexible neck and the articulated human body from low resolution video data remains a hard one. In recent years, integrated face and gait recognition approaches without resorting to 3-D models have achieved some success [1] [2] [3] [4].

Most current gait recognition algorithms rely on the availability of the side view of the subject since human gait or the style of walking is best exposed when one presents a side view to the camera. For face recognition, on the other hand, it is preferred to have frontal views. These conflicting requirements pose some challenges when one attempts to integrate face and gait biometrics in real world applications. In the previous fusion systems [1] [2] [3], the side view of gait and the frontal view of face are used. In [1], Kale et al. present a gait recognition algorithm and a face recognition algorithm based on sequential importance sampling. The database contains video sequences for 30 subjects walking in a single camera scenario. For face recognition, only the final segment of the database presents a nearly frontal view of face and it is used as the probe. The galley consists of static faces for the 30 subjects. Therefore, they perform still-to-video face recognition. In [2] [3], Shakhnarovich et al. compute an image-based visual hull from a set of monocular views of multiple cameras. It is then used to render virtual canonical views for tracking and recognition. They discuss the issues of cross-modal correlation and score transformations for different modalities, and present the cross-modal fusion. In their work, 4 monocular cameras are used to get both the side view of gait and the frontal view of face simultaneously. Recently, Zhou et al. propose a system [4], which combines cues of face profile and gait silhouette from the single camera video sequences. It is based on the fact that a side view of face is more likely to be seen than a frontal view of face when one exposes the best side view of gait to the camera. The data is collected for 14 people with 2 video sequences per person. Even

though face profile in Zhou et al.'s work is used reasonably, it only contains shape information of the side view of face and misses its intensity information. In this paper, an innovative video based fusion system is proposed, aiming at recognizing non-cooperating individuals at a distance in a single camera scenario. Information from two biometric sources, side face and gait, from the single camera video sequence, is combined. Side face, not face profile, includes entire side views of eye, nose and mouth, possessing both shape information and intensity information. Therefore, it has more discriminating power for recognition.

TABLE I

OUR APPROACH FOR INTEGRATING FACE AND GAIT FOR HUMAN RECOGNITION VS. THE PREVIOUS WORK.

Features	[1]	[2] [3]	[4]	This Paper
Biometrics	• Frontal face • Gait	• Frontal face • Gait	• Face profile • Gait	• Side face • Gait
Number of Cameras	1	4	1	1
Face Features and Recognition	• Motion vectors • Time series model • Posterior distribution • MAP	• PCA features of the detected face. • k-NN	• Curvature based features of face profile from the high-resolution image. • Dynamic time warping	• Face features of Enhanced Side Face Image (ESFI) • PCA and MDA combined method • k-NN
Gait Features and Recognition	• Entire canonical view image • Template matching based on dynamic time warping.	• Means and standard deviation of moments, and centroid. • k-NN	• Entire Gait Energy image (GEI) • Template matching	• Gait features of Gait Energy Image (GEI) • PCA and MDA combined method • k-NN
Data	• 30 subjects • Number of sequences per person: not specified. • Static images as the face gallery	• 26 subjects [2] • 2 to 14 sequences per person [2] • 12 subjects [3] • 2 to 6 sequences per person [3]	• 14 subjects • 2 sequences per person	• 45 subjects • 2 to 3 sequences per person
Fusion Methods	• Hierarchical fusion • Sum/Product rule	• Min, Max, Sum and Product rules [2]. • Sum rule [3]	• Hierarchical fusion • Sum and Product rules	• Max, Sum and Product rules
Performance Analysis	• No	• No	• No	• Q statistic

Table I presents a summary of related work and compares it with the work presented in this paper. It is difficult to get reliable information of a side face directly from a video frame for recognition task because of limited resolution. To overcome this problem, we construct Enhanced Side Face Image (ESFI), a higher resolution image compared with the image directly obtained from a single video frame, to fuse information of face from multiple video frames. The idea relies on the fact that the temporally adjacent frames in a video sequence, in which one is walking with a side view to the

camera, contain slightly different, but unique information of a side face. Experiments show that better face features can be extracted from constructed ESFI compared to those from original side face images.

The contributions of this paper are as follows:

• We present a system that integrates side face and gait information from video data. The integration of these two biometric modalities has not been done before.

• Both face and gait recognition systems integrate information over multiple frames in a video sequence for improved performance. High-resolution face images are obtained from video and features from face profile are used for side face normalization.

• The fusion of side face and gait biometrics is done at the match score level by obtaining synthetic match scores and using different fusion schemes. Face features and gait features are obtained separately using PCA and MDA combined method from the Enhanced Side Face Image (ESFI) and the Gait Energy Image (GEI), respectively. The fusion performance is evaluated using the Q statistic.

• Various experiments are performed on 45 people with data from 100 video sequences collected over a period of time. Performance comparisons between different biometrics and different fusion methods are presented.

The paper is organized as follows. Section II presents the overall technical approach. It explains the construction of Enhanced Side Face Image (ESFI) and describes the generation of Gait Energy Image (GEI). It presents PCA and MDA combined method for feature extraction using ESFI and GEI templates. It introduces an approach to generate synthetic match scores for fusion and provides a description of the classification method. In Section III, a number of dynamic video sequences are tested in three experiments using the approach presented. Experimental results are compared and discussed. Finally, Section IV concludes the paper.

II. Technical Approach

The overall technical approach is shown in Figure 1. We first construct Enhanced Side Face Image (ESFI) as the face template and Gait Energy Image (GEI) as the gait template from video sequences. During the training procedure, we perform a component and discriminant analysis separately on face

templates and gait templates obtained from all training videos. As a result, transformation matrices and features that form feature gallery are obtained. During the recognition procedure, each testing video is processed to generate both face templates and gait templates, which are then transformed by the transformation matrices obtained during the training procedure to extract face features and gait features, respectively. These testing features are compared with gallery features in the database, and then different fusion strategies are used to combine the results of face classifier and gait classifier to improve recognition performance.

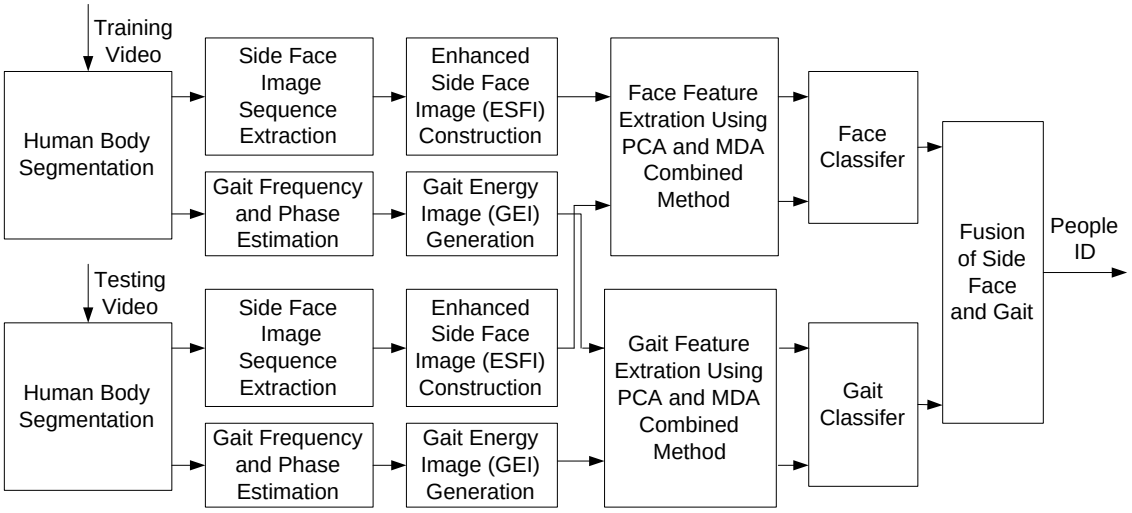


Fig. 1. Technical approach for integrating side face and gait in video.

A. Enhanced Side Face Image (ESFI) Construction

Multiframe resolution enhancement seeks to construct a single high-resolution image from multiple low-resolution images. These low-resolution images must be of the same object, taken from slightly different angles, but not so much as to change the overall appearance of the object in the image.

We use a simple background subtraction method [5] for human body segmentation. A human body is divided into two parts according to the proportion of its parts [6]: from the top of the head to the bottom of the chin, and then from the bottom of the chin to the bottom of the foot. Human head is defined as the part from the top of the head to the bottom of the chin. Considering the height of hair and the length of neck, we assume that the upper 16% of the segmented human body includes the

human head. In this paper, original low-resolution side face images are first localized and extracted by cutting the upper 16% of the segmented human body obtained from multiple video frames.

1) Side Face Image Alignment: Before multiple low-resolution face images can be fused to construct a high-resolution image, motion estimates must be computed to determine pixel displacements between them. It is very important since the quality of a high-resolution image relies on the correctness of low-resolution image alignment. In this paper, the side face images are aligned using a two step procedure. In the first step, an elastic registration algorithm [7] is used for motion estimation in low-resolution side face images. In the second step, a match statistic is introduced to detect and discard images that are poorly aligned. Hence, the quality of constructed high-resolution images can be improved by rejecting such errors.

• Elastic Registration Method: Denote $f(x, y, t)$ and $f(\hat{x}, \hat{y}, t-1)$ as the reference side face image and the image to be aligned, respectively. Assuming that the image intensities are conserved at different times, the motion between images is modeled locally by an affine transform:

$$f(x, y, t) = f(m_1 x + m_2 y + m_5, m_3 x + m_4 y + m_6, t-1)$$

where m_1, m_2, m_3, and m_4 are the linear affine parameters, and m_5 and m_6 are the translation parameters. To account for intensity variations, an explicit change of local contrast and brightness is incorporated into the affine model. Specifically, the initial model takes the form:

$$m_7 f(x, y, t) + m_8 =$$

$$f(m_1 x + m_2 y + m_5, m_3 x + m_4 y + m_6, t-1) \quad (1)$$

where m_7 and m_8 are two new (spatially varying) parameters that embody a change in contrast and brightness, respectively. In order to estimate these parameters, the following quadratic error function is minimized:

$$E(\mathbf{m}) = \sum_{x,y\in\Omega}[m_7 f(x, y, t) + m_8 -$$

$$f(m_1 x + m_2 y + m_5, m_3 x + m_4 y + m_6, t-1)]^2 \quad (2)$$

where $\mathbf{m} = (m_1 m_2 ... m_8)^T$, and Ω denotes a small spatial neighborhood around (x, y). Since this error function is nonlinear in its unknowns, it cannot be minimized analytically. To simplify the minimization, this error function is approximated by using a first-order truncated Taylor series expansion. It now takes the form below.

$$E(\mathbf{m}) = \sum_{x,y \in \Omega} (k - \mathbf{c}^T \mathbf{m})^2 \tag{3}$$

where the scalar k and vector $\mathbf{c}$ are given as:

$$k = f_t - f + x f_x + y f_y \quad \mathbf{c} = (x f_x \quad y f_x \quad x f_y \quad y f_y \quad f_x \quad f_y \quad -f \quad -1)^T \tag{4}$$

where $f_x(\cdot)$, $f_y(\cdot)$ and $f_t(\cdot)$ are the spatial/temporal derivatives of $f(\cdot)$. Minimization of this error function is accomplished by differentiating $E(\mathbf{m})$, setting the result equal to zero and solving for $\mathbf{m}$. The solution is,

$$\mathbf{m} = (\sum_{x,y \in \Omega} \mathbf{c}\mathbf{c}^T)^{-1} (\sum_{x,y \in \Omega} \mathbf{c}k) \tag{5}$$

Intensity variations are typically a significant source of error in differential motion estimation. The addition of the contrast and brightness terms allows us to accurately register images in the presence of local intensity variations. Another important assumption on the model is that the model parameters $\mathbf{m}$ vary smoothly across space. A smoothness constraint on the contrast/brightness parameters has the added benefit of avoiding a degenerate solution where a pure brightness modulation is used to describe the mapping between images.

To begin, the error function is augmented as follows:

$$E(\mathbf{m}) = E_b(\mathbf{m}) + E_s(\mathbf{m}) \tag{6}$$

where $E_b(\mathbf{m})$ is defined without the summation:

$$E_b(\mathbf{m}) = (k - \mathbf{c}^T \mathbf{m})^2 \tag{7}$$

with k and $\mathbf{c}$ as in Equation (4). The new quadratic error term $E_s(\mathbf{m})$ embodies the smoothness constraint:

$$E_s(\mathbf{m}) = \sum_{i=1}^{8} \lambda_i [(\frac{\partial m_i}{\partial x})^2 + (\frac{\partial m_i}{\partial y})^2] \tag{8}$$

where λ_i is a positive constant that controls the relative weight given to the smoothness constraint on parameter m_i. This error function is again minimized by differentiating with respect to the model parameters, setting the result equal to zero and solving $\frac{dE(\mathbf{m})}{d\mathbf{m}} = \frac{dE_b(\mathbf{m})}{d\mathbf{m}} + \frac{dE_s(\mathbf{m})}{d\mathbf{m}} = 0$. Since solving for $\mathbf{m}$ at each pixel location yields an enormous linear system which is intractable to solve, an iterative scheme is used to solve for $\mathbf{m}$ [8]. Now $\mathbf{m}$ is expressed as the following iterative equation:

$$\mathbf{m}^{(j+1)} = (\mathbf{c}\mathbf{c}^T + \mathbf{L})^{-1}(\mathbf{c}k + \mathbf{L}\overline{\mathbf{m}}^{(j)}) \tag{9}$$

where $\overline{\mathbf{m}}$ is the component-wise average of $\mathbf{m}$ over a small spatial neighborhood, and $\mathbf{L}$ is an 8x8 diagonal matrix with diagonal elements λ_i, and zero off the diagonal. On each iteration j, $\overline{\mathbf{m}}^{(j)}$ is estimated from the current $\mathbf{m}^{(j)}$. The initial estimate $\mathbf{m}^{(0)}$ is estimated from Equation (5).

In this paper, a two-level Gaussian pyramid is constructed for both the reference side face image and the side face image to be aligned. The global parameters $\mathbf{m}$ are first estimated at each pyramid level as in Equation (5) for the entire image. Then, the local parameters $\mathbf{m}$ are estimated with $\Omega = 5 \times 5$ as in Equation (5) using the least square algorithm. This estimate of $\mathbf{m}$ is used to bootstrap the iterations in Equation (9). At each iteration, λ_i, $i = 1, ..., 8$, is constant for all $\mathbf{m}$ components and its value is set to 10^{11}. $\overline{m}_i$ is computed by convolving with the 3×3 kernel $(\ 1\ \ 4\ \ 1\ \ ;\ \ 4\ \ 0\ \ 4\ \ ;\ \ 1\ \ 4\ \ 1\ \ )/20$. The number of iteration is 10. This process is repeated at each level of the pyramid. The values of these parameters are chosen empirically and based on the previous motion estimation work [7]. Although the contrast and brightness parameters, m_7 and m_8, are estimated, they are not used when the side face image is aligned to the reference side face image.

• Match Statistic: A match statistic is designed to indicate how well a transformed image aligns with the reference image. It is used to select or reject a low-resolution image during alignment. If the size of the reference image is $M \times N$, the mean square error between the aligned image and the reference image is

$$E = \sum_{x=1}^{M}\sum_{y=1}^{N}[f(x, y, t) - f(m_1x + m_2y + m_5, m_3x + m_4y + m_6, t - 1)]^2/MN.$$

The match statistic of the aligned image is defined as

$$S = 1 - \frac{E}{[\sum_{x=1}^{M} \sum_{y=1}^{N} f^2(x, y, t)]/MN} \tag{10}$$

If the value of S is close to 1, the image at time $t-1$ is well aligned with the image at time t. A very low value indicates misalignment. A perfect match is 1. However, even images that are very well aligned typically do not achieve 1 due to error in the transformation and noise. For improving image quality, the resolution enhancement method discussed next works most effectively when the match values of aligned images are close to 1. A match threshold is specified and any aligned image, whose match statistic falls below the threshold, will not be subsequently used.

The pseudo code for the low-resolution image alignment is shown in Figure 2. Two alignment results with the match statistic S are shown in Figure 3. The reference images and the images to be aligned are from a video sequence, in which a person is walking and exposes a side view to the camera. The reference images in both Figures 3(a) and 3(b) are the same. The time difference between the image to be aligned in Figure 3(a) and the reference image is about 0.033 seconds, and the time difference between the image to be aligned in Figure 3(b) and the reference image is about 0.925 seconds. The S values are 0.95 and 0.86 for Figures 3(a) and 3(b), respectively. Note the differences in the bottom right part of each of the aligned images. We specify the match threshold at 0.9. For 28 out of 100 video sequences used in our experiments, 1 or 2 low-resolution images are discarded from each of the sequences during the image alignment process.

2) Resolution Enhancement Algorithm: An iterative method [9] is used to construct a high-resolution side face image from aligned low-resolution side face images, whose match statistics are above the specified threshold.

• The Imaging Model: The imaging process, yielding the observed side face image sequence f_k, is modeled by:

$$f_k(m, n) = \sigma_k(h(T_k(F(x, y))) + \eta_k(x, y)) \tag{11}$$

where

f_k is the sensed image of the tracked side face in the kth frame.

Align the low-resolution side face image with the reference side face image

Input: *the reference side face image and the side face image to be aligned.*
Output: *the motion vector* $\mathbf{m}$ *and the match statistic* S *of the aligned image.*

1. Global Registration
FOR *each pyramid level from coarse to fine* **DO**
{
Estimate $\mathbf{m}$ *between the newest warped image and the reference image using Equation (5)*
Warp the image to the next level of the pyramid using the newest estimate
}
END FOR
2. Local Registration
FOR *each pyramid level from coarse to fine* **DO**
{
Estimate $\mathbf{m}$ *between the newest warped image and the reference image using Equation (5) with* $\Omega = 5 \times 5$
Warp the image using the newest estimate
FOR *each iteration* **DO**
{
Estimate $\mathbf{m}$ *between the newest warped image and the reference image using Equation (9)*
Warp the image using the newest estimate
}
END FOR
Warp the image to the next level of the pyramid using the newest estimate
}
END FOR
3. Compute the match statistic S *of the aligned image*
4. If $S \geq$ *threshold, keep the low-resolution image; otherwise, discard it*

Fig. 2. Pseudo code for low-resolution image alignment.

(a) A well aligned image with $S = 0.95$.

(b) A bad aligned image with $S = 0.86$.

Fig. 3. Two examples of alignment results with the match statistic S. (a) and (b): the reference image (left), the image to be aligned (middle) and the aligned image (right).

F is a high-resolution image of the tracked side face in a desired reconstruction view. Finding F is the objective of the super-resolution algorithm.

T_k is the 2-D geometric transformation from F to f_k, determined by the 2-D motion parameters $\mathbf{m}$ of the tracked side face in the image plane, which is obtained in Section II-A.1. T_k is assumed to be invertible and does not include the decrease in the sampling rate between F and f_k.

h is a blurring operator, determined by the Point Spread Function (PSF) of the sensor. We use a circular averaging filter with radius 2 as PSF.

η_k is an additive noise term.

σ_k is a down sampling operator which digitizes and decimates the image into pixels and quantizes the resulting pixel values.

The receptive field (in F) of a detector whose output is the pixel $f_k(m,n)$ is uniquely defined by its center (x,y) and its shape. The shape is determined by the region of the blurring operator h, and by the inverse geometric transformation T_k^{-1}. Similarly, the center (x,y) is obtained by $T_k^{-1}(m,n)$. The resolution enhancement algorithm aims to construct a higher resolution image $\hat{F}$, which approximates F as accurately as possible, and surpasses the visual quality of the observed images in $\{f_k\}$.

• Algorithm for Resolution Enhancement: The algorithm for creating higher resolution images is iterative. Starting with an initial guess $F^{(0)}$ for the high-resolution side face image, the imaging process is simulated to obtain a set of low-resolution side face images $\{f_k^{(0)}\}_{k=1}^K$ corresponding to the observed input images $\{f_k\}_{k=1}^K$. If $F^{(0)}$ were the correct high-resolution side face image, then the simulated images $\{f_k^{(0)}\}_{k=1}^K$ should be identical to the observed low-resolution side face image $\{f_k\}_{k=1}^K$. The difference images $\{f_k - f_k^{(0)}\}_{k=1}^K$ are used to improve the initial guess by "back projecting" each value in the difference images onto its receptive field in $F^{(0)}$, yielding an improved high-resolution side face image $F^{(1)}$. This process is repeated iteratively to minimize the error function:

$$e^{(n)} = \sqrt{\frac{1}{K}\sum_{k=1}^{K} \|f_k - f_k^{(n)}\|^2} \tag{12}$$

The imaging process of f_k at the nth iteration is simulated by:

$$f_k^{(n)} = (T_k(F^{(n)}) * h) \downarrow s \tag{13}$$

where $\downarrow s$ denotes a down sampling operator by a factor s, and $*$ is the convolution operator. The iterative update scheme of the high-resolution image is expressed by:

$$F^{(n+1)} = F^{(n)} + \frac{1}{K}\sum_{k=1}^{K} T_k^{-1}(((f_k - f_k^{(n)}) \uparrow s) * p) \tag{14}$$

where K is the number of low-resolution side face images. $\uparrow s$ is an up sampling operator by a factor s, and p is a "back projection" kernel, determined by h. T_k is 2-D motion parameters. The averaging process reduces additive noise.

Construct the high-resolution side face image from the low-resolution side face images

Input: *the observed input images* $\{f_k\}_{k=1}^K$ *and the corresponding motion vectors* $\{\mathbf{m}_k\}_{k=1}^K$
Output: *the high-resolution image* F.

1. Start with iteration $n = 0$
2. Obtain an initial guess $F^{(0)}$ *for the high-resolution image using bilinear interpolation*
3. Obtain a set of low-resolution images $\{f_k^{(n)}\}_{k=1}^K$ *using Equation (13)*
4. Obtain an improved high-resolution image $F^{(n+1)}$ *using Equation (14)*
5. Let $n = n + 1$
6. If $n <= N$*, go to step 3; otherwise, stop*

Fig. 4. Pseudo code for high-resolution image construction.

In this paper, we use a sampling factor $s = 2$. An initial guess $F^{(0)}$ for the high resolution image is obtained by up sampling a low-resolution image using bilinear interpolation. Ten low-resolution side face images contribute to a high-resolution side face image. The high-resolution image is obtained after 10 iterations ($N = 10$) .

The pseudo code for the high-resolution image construction is shown in Figure 4. Figure 5 shows four examples of low-resolution face images and reconstructed high-resolution face images. The resolution of the low-resolution side face images is 68×68 and the resolution of the high-resolution side face images is 136×136. For comparison, we resize the low-resolution face images using bilinear interpolation. From this figure, we can see that the quality of the reconstructed high-resolution images is much better than the resized low-resolution images.

Fig. 5. Four examples of resized low-resolution face images (top) and constructed high-resolution face images (bottom).

3) Side Face Normalization: Before feature extraction, all high-resolution side face images are normalized. The normalization is based on the locations of nasion, pronasale and throat on the

face profile. These three fiducial points are identified by using a curvature based fiducial extraction method [10]. It is explained as follows.

We apply a canny edge detector to the side face image. After edge linking and thinning, the profile of a side face is extracted as the leftmost points different from background, which contain fiducial points like nasion, pronasale, chin and throat. The profile consists of a set of points $T = (x, y)$, where x is a row index and y is a column index of a pixel. Then, a Gaussian scale-space filter is applied to this $1D$ curve to reduce noise. The convolution between Gaussian kernel $g(x, \sigma)$ and signal $f(x)$ depends both on x, the signal's independent variable, and on σ, the Gaussian's standard deviation. It is given by

$$F(x, \sigma) = f(x) \oplus g(x, \sigma) = \int_{-\infty}^{\infty} f(u) \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-u)^2}{2\sigma^2}} du \tag{15}$$

where $\oplus$ denotes convolution with respect to x. The bigger the σ, the smoother the $F(x, \sigma)$. The curve T is parameterized as $T(u) = (x(u), y(u))$ by the arc length parameter u. An evolved version of T is $T_\sigma(u) = (X(u, \sigma), Y(u, \sigma))$, where $X(u, \sigma) = x(u) \oplus g(u, \sigma)$ and $Y(u, \sigma) = y(u) \oplus g(u, \sigma)$.

Curvature κ on T_σ is computed as:

$$\kappa(u, \sigma) = \frac{X_u(u, \sigma)Y_{uu}(u, \sigma) - X_{uu}(u, \sigma)Y_u(u, \sigma)}{\left(X_u(u, \sigma)^2 + Y_u(u, \sigma)^2\right)^{1.5}} \tag{16}$$

where the first and second derivatives of X and Y can be computed as:

$$X_u(u, \sigma) = x(u) \oplus g_u(u, \sigma) \quad X_{uu}(u, \sigma) = x(u) \oplus g_{uu}(u, \sigma)$$

$$Y_u(u, \sigma) = y(u) \oplus g_u(u, \sigma) \quad Y_{uu}(u, \sigma) = y(u) \oplus g_{uu}(u, \sigma)$$

$g_u(u, \sigma)$ and $g_{uu}(u, \sigma)$ are the first derivative and the second derivative of Gaussian Kernel.

To localize the fiducial points, the curvature of a profile is first computed at an initial scale and the locations, where the local maxima of the absolute values occur, are chosen as corner candidates. These locations are tracked down and the fiducial points are identified at lower scales. The initial scale must be large enough to remove noise and small enough to retain the real corners. Our method has advantages in that it does not depend on too many parameters and not require any thresholds. It is also fast and simple. The complete process to find the fiducial points is described as follows:

Step 1: Compute the curvature of a profile at an initial scale, find all points with the large absolute curvature values as corner candidates and track them down to lower scales.

Step 2: Regard the rightmost point in the candidate set as the throat.

Step 3: Regard the pronasale as one of the two leftmost candidate points in the middle part of the profile and then identify it using the curvature value around this point.

Step 4: Assume that there are no candidate points between pronasale and nasion and identify the first candidate point above the pronasale as nasion.

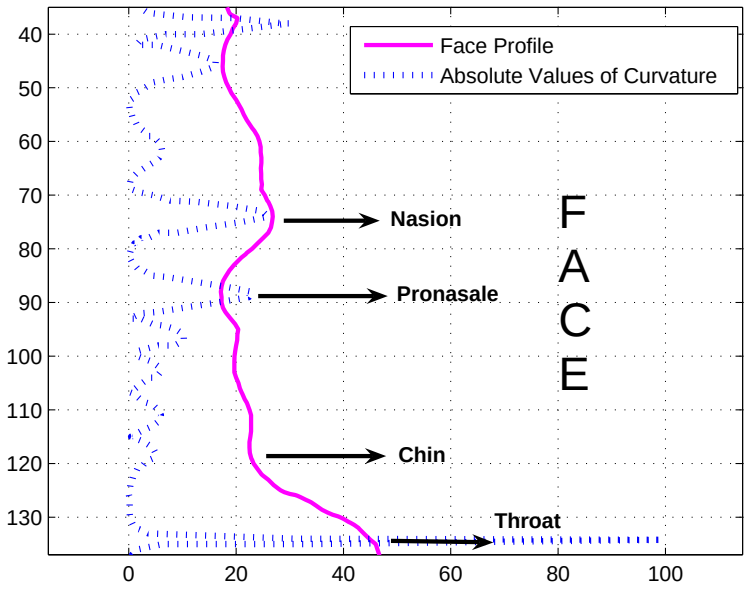
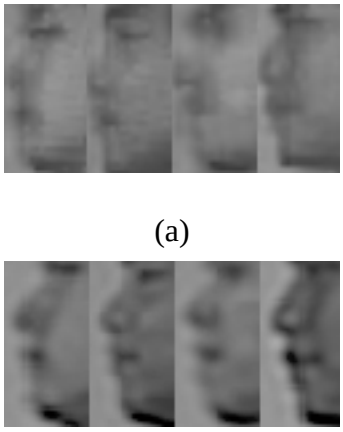


Fig. 6. The extracted face profile and the absolute values of curvature.

Figure 6 shows the extracted face profile and the absolute values of curvature. We amplify the absolute values of curvature 20 times in order to show them more clearly. It is clear that the locations of the fiducial points, including nasion, pronasale and throat, have large curvature values. Given a set of high-resolution images and the three fiducial points of each face image, affine transformations are computed between the first image and all the other images. Subsequently, images are cropped as follows: the highest point is defined as the point six pixels above nasion; the lowest point is defined as the throat; the leftmost point is defined as the point 4 pixels to the left of pronasion; and the rightmost point is defined as the one, which is half of the height of the cropped image and is to the right of the leftmost point. All cropped images are further normalized to the size of 64×32. We call

these images as Enhanced Side Face Images (ESFIs). Similarly, Original Side Face Image (OSFI) is a subimage from the normalized version of the low-resolution side face image. It is obtained by the similar process explained above. The size of OSFI is 34×18. Examples of resized OSFIs and ESFIs for four people are shown for comparison in Figure 7. Clearly, ESFIs have better quality than OSFIs.



(a)

(b)

Fig. 7. Examples of 4 people: (a) Resized OSFIs (b) ESFIs.

B. Gait Energy Image (GEI) Construction

Regular human walking can be considered as cyclic motion where human motion repeats at a stable frequency [11]. Therefore, it is possible to divide the entire gait sequence into cycles. Since human body segmentation is performed on the original human walking sequences, we begin with the extracted binary silhouette image sequences. The silhouette preprocessing includes size normalization (proportionally resizing each silhouette image so that all silhouettes have the same height) and horizontal alignment (centering the upper half silhouette part with respect to its horizontal centroid). In a preprocessed silhouette sequence, the time series signal of lower half silhouette size from each frame indicates the gait frequency and phase information. We estimate the gait frequency and phase by maximum entropy spectrum estimation [12] from the time series signal.

Given the preprocessed binary gait silhouette image $B_t(x, y)$ at time t in a sequence, the grey-level

gait energy image (GEI) is defined as follows [11]:

$$G(x,y) = \frac{1}{N}\sum_{t=1}^{N} B_t(x,y) \tag{17}$$

where N is the number of frames in the complete cycle(s) of a silhouette sequence, t is the frame

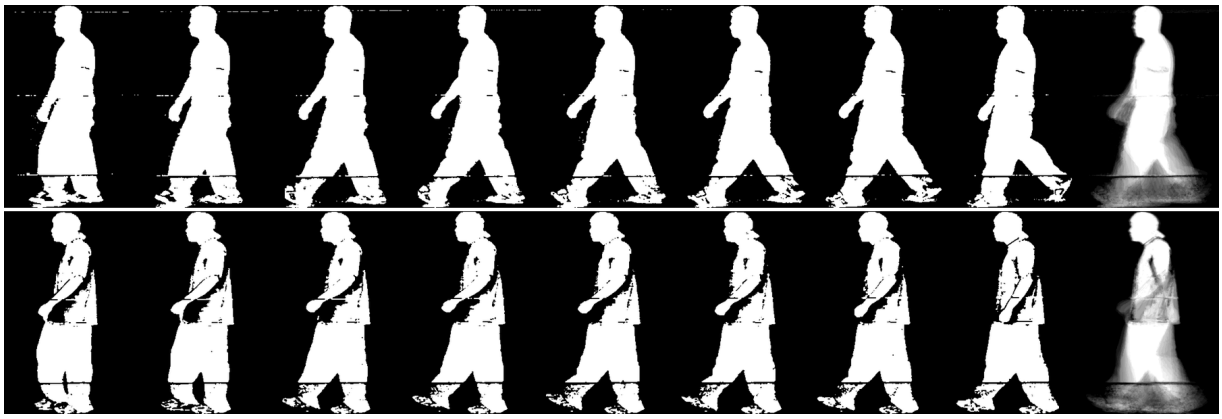


Fig. 8. Two examples of normalized and aligned silhouette images in a gait cycle. The right most images are the corresponding gait energy images (GEIs).

number of the sequence (moment of time), and x and y are values in the 2D image coordinate. Figure 8 shows the sample silhouette images in a gait cycle from 2 people and the right most images are the corresponding GEIs. As expected, GEI reflects major shapes of silhouettes and their changes over the gait cycle. It accounts for human walking at different speeds. It is referred as the gait energy image because: (a) each silhouette image is the space-normalized energy image of human walking at this moment; (b) GEI is the time-normalized accumulative energy image of human walking in the complete cycle(s); (c) a pixel with higher intensity value in GEI means that human walking occurs more frequently at this position (i.e., with higher energy). GEI has several advantages over the gait representation of binary silhouette sequence. GEI is not sensitive to incidental silhouette errors in individual frames. Moreover, with such a 2D template, we do not need to consider the time moment of each frame, and the incurred errors can be, therefore, avoided.

C. Human Recognition Using ESFI and GEI

1) Feature Learning Using PCA and MDA Combined Method: In this paper, PCA and MDA combined method [13] is applied to face templates, ESFIs, and gait templates, GEIs, separately to get

low dimensional feature representation for side face and gait. PCA reduces the dimension of feature space, and MDA automatically identifies the most discriminating features.

Let $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, $\mathbf{x}_k \in \mathbb{R}^N$, be n random vectors representing n ESFIs or n GEIs, where N is the dimensionality of the image. The covariance matrix is defined as $\mathbf{\Sigma_x} = E([\mathbf{x} - E(\mathbf{x})][\mathbf{x} - E(\mathbf{x})]^T)$, where $E(\cdot)$ is the expectation operator and T denotes the transpose operation. The covariance matrix $\mathbf{\Sigma_x}$ can be factorized into the following form:

$$\mathbf{\Sigma_x} = \mathbf{\Phi \Lambda \Phi} \tag{18}$$

where $\mathbf{\Phi} = [\mathbf{\Phi}_1 \mathbf{\Phi}_2 ... \mathbf{\Phi}_N] \in \mathbb{R}^{N \times N}$ is the orthogonal eigenvector matrix of $\mathbf{\Sigma_x}$; $\mathbf{\Lambda} = \{\mathbf{\Lambda}_1 \mathbf{\Lambda}_2 ... \mathbf{\Lambda}_N\} \in \mathbb{R}^{N \times N}$ is the diagonal eigenvalue matrix of $\mathbf{\Sigma_x}$ with diagonal elements in descending order. One important property of PCA is its optimal signal reconstruction in the sense of minimum mean square error (MSE) when only a subset of principal components are used to represent the original signal. An immediate application of this property is the dimensionality reduction:

$$\mathbf{y}_k = \mathbf{P}_{pca}^T [\mathbf{x}_k - E(\mathbf{x})] \quad k = 1, ..., n. \tag{19}$$

where $\mathbf{P}_{pca} = [\mathbf{\Phi}_1 \mathbf{\Phi}_2 ... \mathbf{\Phi}_m], m < N$. The lower dimensional vector $\mathbf{y}_k \in \mathbb{R}^m$ captures the most expressive features of the original data $\mathbf{x}_k$.

MDA seeks a transformation matrix $\mathbf{W}$ that maximizes the ratio of the between-class scatter matrix $\mathbf{S}_B$ to the within-class scatter matrix $\mathbf{S}_W$: $J(\mathbf{W}) = \frac{|\mathbf{W}^T \mathbf{S}_B \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_W \mathbf{W}|}$. Suppose that $\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_c$ and $n_1, n_2, ..., n_c$ denote the classes and the number of images within each class, respectively, with $n = n_1 + n_2 + ... + n_c$ and $\mathbf{w} = \mathbf{w}_1 \cup \mathbf{w}_2 \cup ... \cup \mathbf{w}_c$. c is the number of classes. The within-class scatter matrix is $\mathbf{S}_W = \sum_{i=1}^{c} \sum_{\mathbf{y} \in \mathbf{w}_i} (\mathbf{y} - \mathbf{M}_i)(\mathbf{y} - \mathbf{M}_i)^T$ and the between-class scatter matrix is $\mathbf{S}_B = \sum_{i=1}^{c} n_i (\mathbf{M}_i - \mathbf{M})(\mathbf{M}_i - \mathbf{M})^T$, where $\mathbf{M}_i = \frac{1}{n_i} \sum_{\mathbf{y} \in \mathbf{w}_i} \mathbf{y}$ and $\mathbf{M} = \frac{1}{n} \sum_{\mathbf{y} \in \mathbf{w}} \mathbf{y}$ are the means of the class i and the grand mean, respectively. $J(\mathbf{W})$ is maximized when the columns of $\mathbf{W}$ are the generalized eigenvectors of $\mathbf{S}_B$ and $\mathbf{S}_W$ corresponding to the largest generalized eigenvalues in

$$\mathbf{S}_B \mathbf{\Psi}_i = \lambda_i \mathbf{S}_W \mathbf{\Psi}_i \tag{20}$$

There are no more than $c - 1$ nonzero eigenvalues λ_i and the corresponding eigenvectors $\mathbf{\Psi}_i$. The

transformed feature vector is obtained as follows:

$$\mathbf{z}_k = \mathbf{P}_{mda}^T \mathbf{y}_k = \mathbf{P}_{mda}^T \mathbf{P}_{pca}^T [\mathbf{x}_k - E(\mathbf{x})] = \mathbf{Q}[\mathbf{x}_k - E(\mathbf{x})] \quad k = 1, ..., n. \tag{21}$$

where $\mathbf{P}_{mda} = [\mathbf{\Psi}_1 \mathbf{\Psi}_2 ... \mathbf{\Psi}_k], k < c$ and $\mathbf{Q}$ is the overall transformation matrix. We can choose k to perform feature selection and dimensionality reduction. The choice of the range of PCA and the dimension of MDA reflects both the energy need and the magnitude requirement. The lower dimensional vector $\mathbf{z}_k \in \mathbb{R}^k$ captures the most expressive and discriminating features of the original data $\mathbf{x}_k$.

2) Recognition by Integrating ESFI and GEI: We train face templates and gait templates separately for feature extraction. Let $\{\mathbf{F}\}$ be the set of all training face templates, and $\mathbf{Q}^f$ be the corresponding face transformation matrix. Let $\{\mathbf{G}\}$ be the set of all training gait templates, and $\mathbf{Q}^g$ be the corresponding gait transformation matrix. Let $\{\mathbf{f}_i\}$ be the set of face feature vectors belonging to the ith class, and $\{\mathbf{g}_i\}$ be the set of gait feature vectors belonging to the ith class, $i = 1, 2, ..., c$, where c is the number of classes in the gallery. Given a testing video P, we follow the procedure explained in Section II-A and Section II-B to generate the set of testing face templates $\{\hat{\mathbf{F}}_P\}$ and the set of testing gait templates $\{\hat{\mathbf{G}}_P\}$, respectively. The corresponding face and gait feature vector sets are obtained using Equation (21) as follows:

$$\begin{aligned} \{\hat{\mathbf{f}}_P\} &: \hat{\mathbf{f}}_{Pj} = \mathbf{Q}^f \hat{\mathbf{F}}_{Pj} \quad j = 1, 2, ..., n_f \\ \{\hat{\mathbf{g}}_P\} &: \hat{\mathbf{g}}_{Pj} = \mathbf{Q}^g \hat{\mathbf{G}}_{Pj} \quad j = 1, 2, ..., n_g \end{aligned} \tag{22}$$

where n_f is the number of testing face templates and n_g is the number of testing gait templates.

The Euclidean distance is used as the similarity measure for the face classifier and the gait classifier. From the classifier based on face templates, we obtain

$$D(\hat{\mathbf{f}}_{Pj}, \mathbf{f}_i) = ||\hat{\mathbf{f}}_{Pj} - \mathbf{m}_{fi}|| \quad i = 1, 2, ..., c \quad j = 1, 2, ..., n_f \tag{23}$$

where $\mathbf{m}_{fi} = \frac{1}{N_{fi}} \sum_{\mathbf{f} \in \mathbf{f}_i} \mathbf{f}$, $i = 1, 2, ..., c$, is the prototype of class i for face and N_{fi} is the number of face feature vectors in $\{\mathbf{f}_i\}$. We assign the testing video P to class k if

$$D(\hat{\mathbf{f}}_P, \mathbf{f}_k) = \min_{i=1}^{c} \min_{j=1}^{n_f} D(\hat{\mathbf{f}}_{Pj}, \mathbf{f}_i) \tag{24}$$

From the classifier based on gait templates, we obtain

$$D(\hat{\mathbf{g}}_{Pj}, \mathbf{g}_i) = ||\hat{\mathbf{g}}_{Pj} - \mathbf{m}_{gi}|| \quad i = 1, 2, ..., c \quad j = 1, 2, ..., n_g \tag{25}$$

where $\mathbf{m}_{gi} = \frac{1}{N_{gi}} \sum_{\mathbf{g} \in \mathbf{g}_i} \mathbf{g}$, $i = 1, 2, ..., c$, is the prototype of class i for gait and N_{gi} is the number of gait feature vectors in $\{\mathbf{g}_i\}$. We assign the testing video P to class k if

$$D(\hat{\mathbf{g}}_P, \mathbf{g}_k) = \min_{i=1}^{c} \min_{j=1}^{n_g} D(\hat{\mathbf{g}}_{Pj}, \mathbf{g}_i) \tag{26}$$

Before combination of the results of face classifier and the results of gait classifier, it is necessary to map distances obtained from the different classifiers to the same range of values. We use exponential transformation here. Given that the distance for a probe X are $S_1, S_2, ..., S_c$, we obtain the normalized match scores as

$$S'_i = \frac{exp(-S_i)}{\sum_{i=1}^{c} exp(-S_i)} \quad i = 1, 2, ..., c \tag{27}$$

After normalization, the match scores of face templates and the match scores of gait templates from the same class are fused using different fusion methods. Since face and gait can be regraded as two independent biometrics in our scenario, synchronization is totally unnecessary for them. To take advantage of information for a walking person in video, we use all the possible combinations of face match scores and gait match scores to generate new match scores, which encode information from both face and gait. The new match scores are called *synthetic match scores*, defined as

$$D_t(\{\hat{\mathbf{f}}_P, \hat{\mathbf{g}}_P\}, \{\mathbf{f}_l, \mathbf{g}_l\}) = R\{D'(\hat{\mathbf{f}}_{Pi}, \mathbf{f}_l), D'(\hat{\mathbf{g}}_{Pj}, \mathbf{g}_l)\}$$

$$i = 1, 2, ..., n_f \quad j = 1, 2, ..., n_g \quad t = 1, 2, ..., n_f n_g \quad l = 1, 2, ..., c \tag{28}$$

where D' means the normalized match score of the corresponding distance D, and $R\{,\}$ means a fusion method. In this paper, we use Sum, Product and Max rules. It is reasonable to generate synthetic match scores using Equation (28), since ESFI is built from multiple video frames and GEI is a compact spatio-temporal representation of gait in video. In this paper, we use 2 face match scores and 2 gait match scores to generate 4 synthetic match scores for one person from each video.

Distances representing dissimilarity become match scores representing similarity by using Equation (27), so the unknown person should be classified to the class for which the synthetic match score is the largest. We assign the testing video P to class k if

$$D(\{\hat{\mathbf{f}}_P, \hat{\mathbf{g}}_P\}, \{\mathbf{f}_k, \mathbf{g}_k\}) = \max_{l=1}^{c} \max_{t=1}^{n_f n_g} D_t(\{\hat{\mathbf{f}}_P, \hat{\mathbf{g}}_P\}, \{\mathbf{f}_l, \mathbf{g}_l\}) \tag{29}$$

Since we obtain more than one synthetic match scores after fusion for one testing video sequence, Equation (29) means the unknown person is classified to the class which gets the maximum synthetic match score out of all the synthetic match scores corresponding to all the classes.

III. Experimental Results

A. Experiments and Parameters

We perform three experiments to test our approach. The data are obtained by a Sony DCR-VX1000 digital video camera recorder operating at 30 frames per second. We collect video sequences of 45 people, who are walking in outdoor condition and expose a side view to the camera. The number of sequences per person varies from 2 to 3. The resolution of each frame is 720x480. The distance between people and the video camera is about 10 feet. Each video sequence includes only one person.

In Experiment 1, the data consists of 90 video sequences of 45 people. Each person has two video sequences, one for training and the other one for testing. For the same person, the clothes are the same in the training sequence and the testing sequence. In Experiment 2, the data consists of 90 video sequences of 45 people. Each person has two video sequences, one for training and the other one for testing. For 10 of 45 people, the clothes are different in the training sequences and the testing sequences, and the data are collected on two separate days about 1 month apart. For the other 35 people, the clothes are the same in the training sequences and the testing sequences. In Experiment 3, we use the same data as in Experiment 2. The difference between them is that we use different number of ESFIs and GEIs in the testing procedure. Table II summaries the key features of the three experiments.

For gait, we obtain 2 complete walking cycles from a video sequence according to the gait frequency and gait phase. Each walking cycle includes about 20 frames. We construct 2 GEIs corresponding

TABLE II

SUMMARY OF THREE EXPERIMENTS.

Data	Experiments		
	1	2	3
Number of subjects	45	45	45
Number of subjects with changed clothes	0	10	10
Number of GEIs for testing per video	2	2	1 or 2
Number of ESFIs for testing per video	2	2	1 or 2

to 2 walking cycles from one video sequence. The resolution of each GEI is 300x200. For face, we also construct 2 high-resolution side face images from one video sequence. The match threshold (the match statistic S) for aligned low-resolution side face images is specified at 0.9. Each high-resolution side face image is built from 10 low-resolution side face images that are extracted from adjacent video frames. The resolution of low-resolution side face images is 68x68 and the resolution of reconstructed high-resolution side face images is 136x136. After normalization (see Section II-A.3), the resolution of ESFI is 64x32. Recognition performance is used to evaluate our method in the three experiments. For a video sequence, it is defined as the ratio of the number of the correctly recognized people to the number of all the people. To analyze the performance of our method more insightfully, we provide the error index that gives the numbers of misclassified sequences. For comparison, we also show the performance using face features from the Original Side Face Images (OSFIs) to demonstrate the performance improvement by using constructed ESFIs. The resolution of OSFI is 34x18. The procedures of feature extraction, synthetic match score generation and classification are the same for ESFI and OSFI.

1) Experiment 1: Figure 9 and Figure 10 show the data used in Experiment 1. We name 45 people from 1 to 45 and each person has 2 video sequences. For each of the 45 people, some frames of the training sequence and the testing sequence are shown. Since we construct 2 GEIs and 2 ESFIs for each sequence, we totally obtain 90 ESFIs and 90 GEIs as the gallery and another 90 ESFIs and 90 GEIs as the probe. After fusion, as explained in Section II-C.2, 4 synthetic match scores are generated based on 2 face match scores and 2 gait match scores for one person from each video. Totally, we

have 180 synthetic match scores corresponding to 45 people in the gallery and 180 synthetic match scores corresponding to 45 people in the probe. Table III shows the performance of single biometric. Table IV shows the performance of fusion using different combination rules. In Table III and Table IV, the error index gives the number of misclassified sequence.

TABLE III

EXPERIMENT 1: SINGLE BIOMETRIC PERFORMANCE AND ERROR INDEX OF INDIVIDUALS.

Performance	Biometric		
	Original Face (OSFI)	Enhanced Face (ESFI)	Gait (GEI)
Recognition Rate	73.3%	91.1%	93.3%
Error Index	1, 6, 10, 12, 14, 18, 20, 22, 26, 28, 42, 43	13, 16, 21, 35	4, 15, 26

TABLE IV

EXPERIMENT 1: FUSED BIOMETRIC PERFORMANCE AND ERROR INDEX OF INDIVIDUALS.

Fusion Method		Sum Rule	Product Rule	Max Rule
OSFI & GEI	Recognition Rate	93.3%	95.6%	93.3%
	Error Index	4, 10, 26	4, 26	4, 10, 26
ESFI & GEI	Recognition Rate	95.6%	95.6%	97.8%
	Error Index	4, 26	4, 26	26

From Table III, we can see that 73.3% people are correctly recognized by OSFI (12 errors out of 45 people), 91.1% people are correctly recognized by ESFI (4 errors out of 45 people) and 93.3% people are correctly recognized by GEI (3 errors out of 45 people). Among performance of fusion methods in Table IV, Max rule based on ESFI and GEI performs the best at the recognition rate of 97.8% (1 errors out of 45 people), followed by Sum rule and Product rule at 95.6% (2 errors out of 45 people). For fusion based on OSFI and GEI, the best performance is achieved by Product rule at 95.6%, followed by Sum rule and Product rule at 93.3%. It is clear that fusion based on ESFI and GEI always has better performance than fusion based on OSFI and GEI, except using Product rule where they are the same. Figure 11 shows people (video sequences) misclassified by integrating ESFI and GEI using different fusion rules. It is clear that the only person (26), who is misclassified by the

Fig. 9. Data in Experiment 1: Video sequences from number 1 to 23.

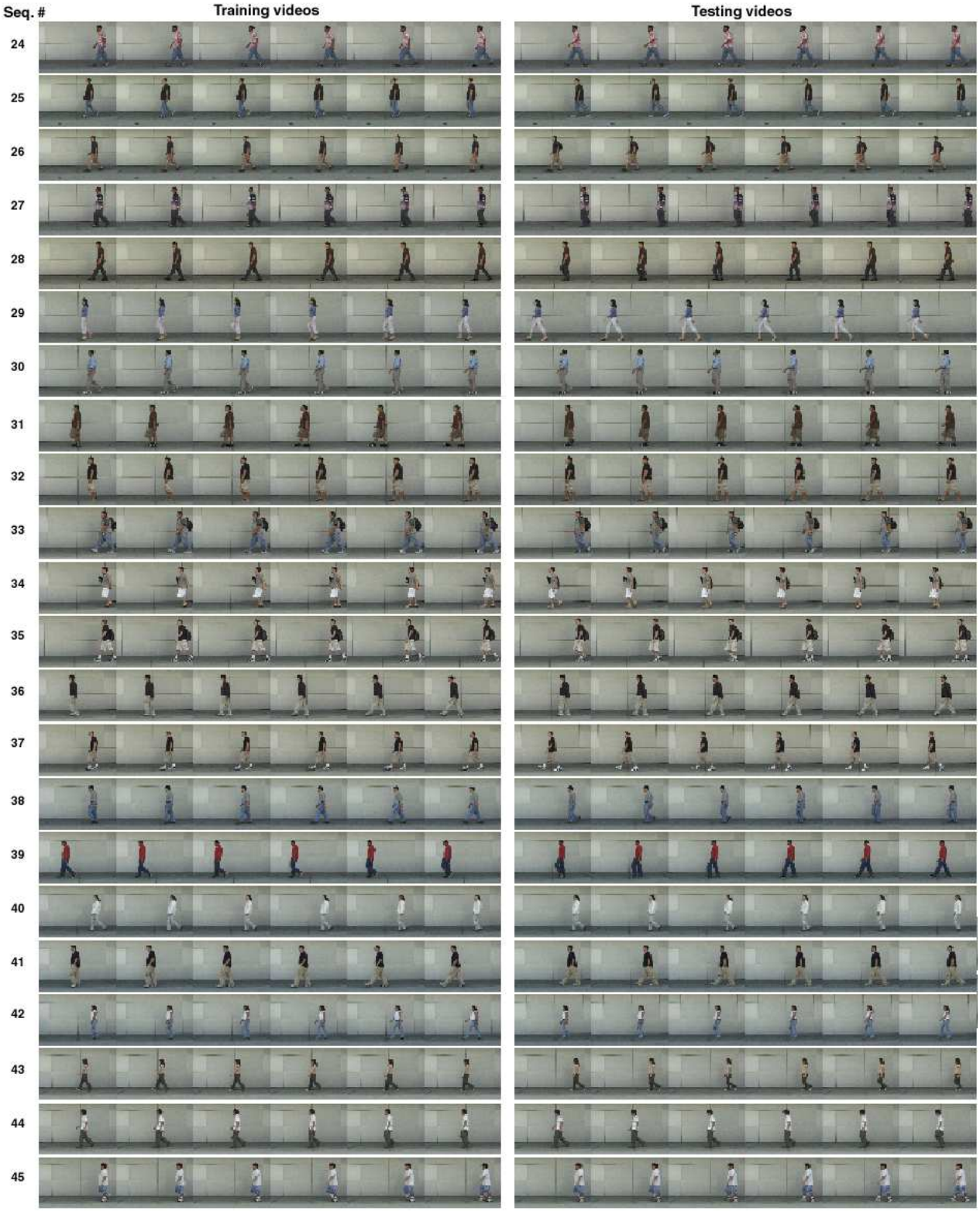
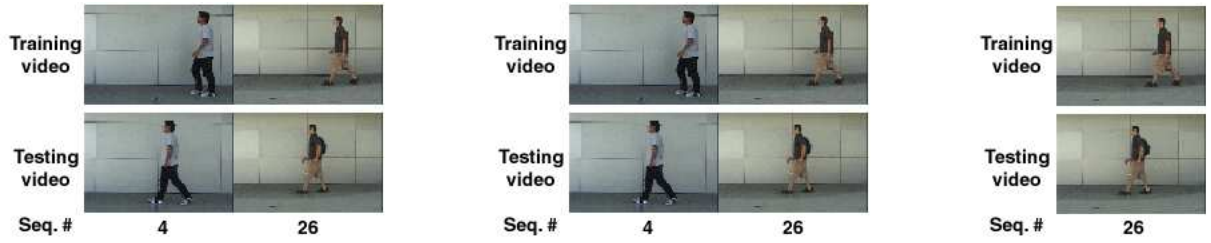


Fig. 10. Data in Experiment 1: Video sequences from number 24 to 45.



(a) Errors by Sum rule. (b) Errors by Product rule. (c) Errors by Max rule.

Fig. 11. Experiment 1: People misclassified by the integrated classifier based on ESFI and GEI using different fusion rules (see Table IV). For each person, one frame of the training video sequence and one frame of the testing video sequence are shown for comparison.

Max rule, has a backpack in the testing sequence that does not occur in the training sequence. This difference makes both the gait classifier and the fused classifier unable to recognize him.

2) Experiment 2: The data used in Experiment 2 are obtained by substituting 10 testing video sequences of Experiment 1 with the other 10 testing video sequences shown in Figure 12. We use the same order as in Experiment 1 to name 45 people. Compared with the data in Experiment 1, the 10 replaced testing video sequences are $\{1, 2, 5, 6, 8, 9, 10, 13, 19, 40\}$. Therefore, 10 out of 45 people in Experiment 2 wear different clothes in the training sequences and the testing sequences, and for each of the 10 people, two video sequences are collected on two separate days about 1 month apart. We construct 2 GEIs and 2 ESFIs from each sequence, so we totally obtain 90 ESFIs and 90 GEIs as the gallery and another 90 ESFIs and 90 GEIs as the probe for 45 people. After fusion, as explained in Section II-C.2, we have 180 synthetic match scores corresponding to 45 people in the gallery and 180 synthetic match scores corresponding to 45 people in the probe. Table V shows the performance of individual biometric. Table VI shows the performance of fusion using different combination rules. In Table V and Table VI, the error index gives the number of misclassified sequence.

TABLE V

EXPERIMENT 2: SINGLE BIOMETRIC PERFORMANCE AND ERROR INDEX OF INDIVIDUALS.

Performance	Biometric		
	Original Face (OSFI)	Enhanced Face (ESFI)	Gait (GEI)
Recognition Rate	64.4%	80%	82.2%
Error Index	1, 2, 5, 6, 8, 9, 13, 18, 19, 20, 26, 28, 34, 40, 42, 43	1, 2, 5, 8, 11, 13, 30, 35, 42	2, 5, 6, 8, 13, 19, 26, 40

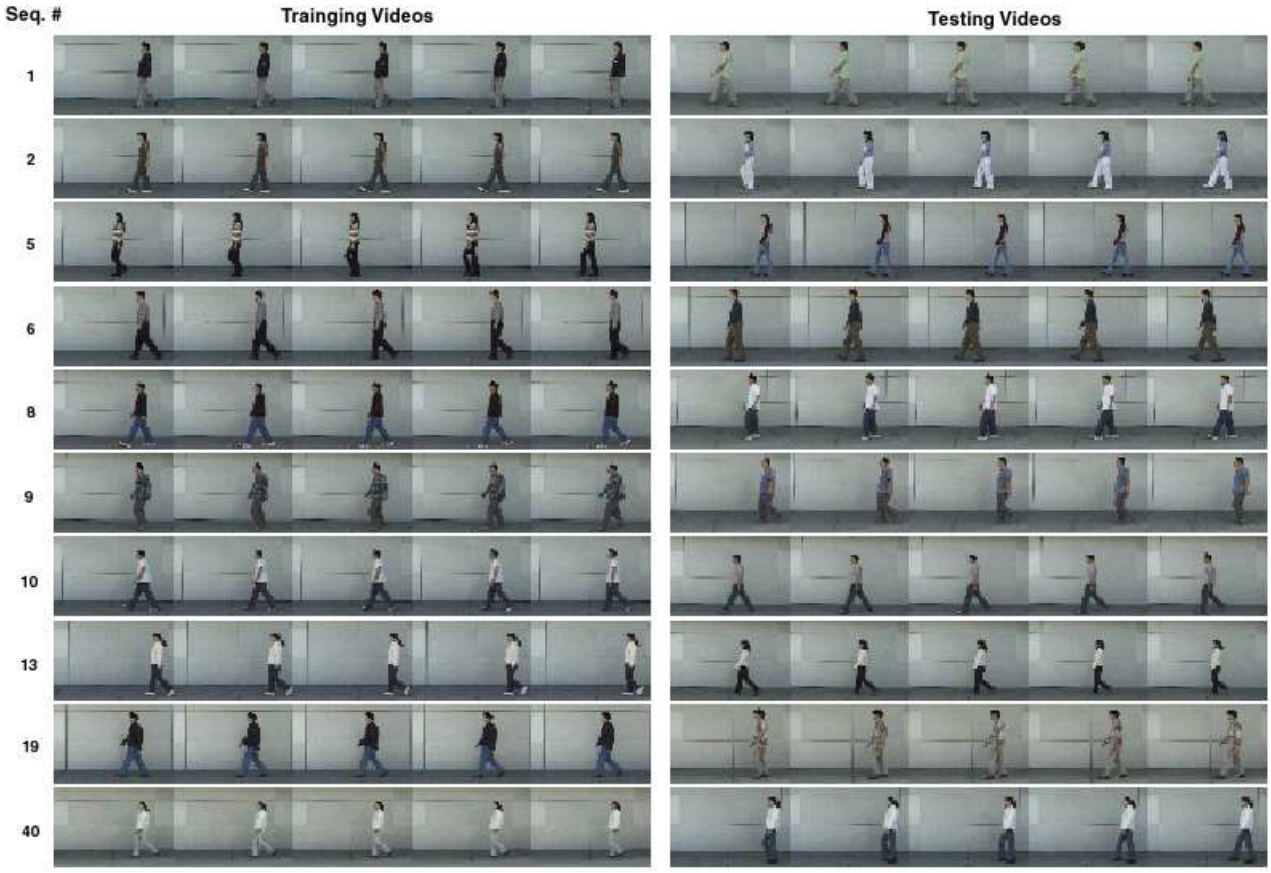


Fig. 12. Data in Experiment 2: 10 updated video sequences $\{1, 2, 5, 6, 8, 9, 10, 13, 19, 40\}$.

TABLE VI

EXPERIMENT 2: FUSED BIOMETRIC PERFORMANCE AND ERROR INDEX OF INDIVIDUALS.

Fusion Method		Sum Rule	Product Rule	Max Rule
OSFI & GEI	Recognition Rate	82.2%	82.2%	82.2%
	Error Index	2, 5, 6, 8, 13, 19, 26, 40	2, 5, 6, 8, 13, 19, 26, 40	2, 5, 6, 8, 13, 19, 26, 40
ESFI & GEI	Recognition Rate	88.9%	82.2%	88.9%
	Error Index	2, 5, 6, 8, 13	2, 5, 6, 8, 13, 19, 26, 40	2, 5, 6, 8, 13

From Table V, we can see that 64.4% people are correctly recognized by OSFI (16 errors out of 45 people), 80% people are correctly recognized by ESFI (9 errors out of 45 people) and 82.2% people are correctly recognized by GEI (8 errors out of 45 people). Compared with the performance of individual biometric in Experiment 1 in Table III, all the performance of individual biometric in Experiment 2 decreases to some extent. It is reasonable since gait recognition based on GEI is not only affected by the walking style of a person, but also by the shape of a human body. Changing

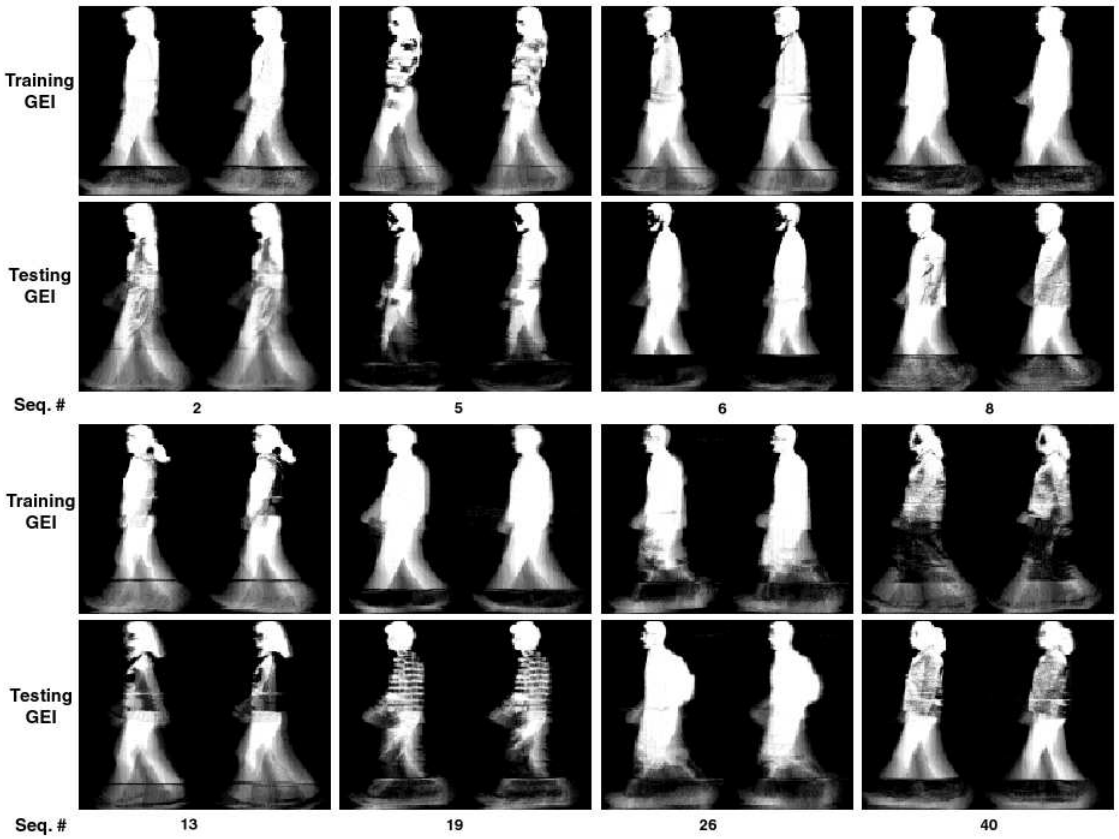


Fig. 13. Experiment 2: GEIs of people misclassified by the gait classifier (see Table V). For each person, 2 GEIs of the training video sequence and 2 GEIs of the testing video sequence are shown for comparison.

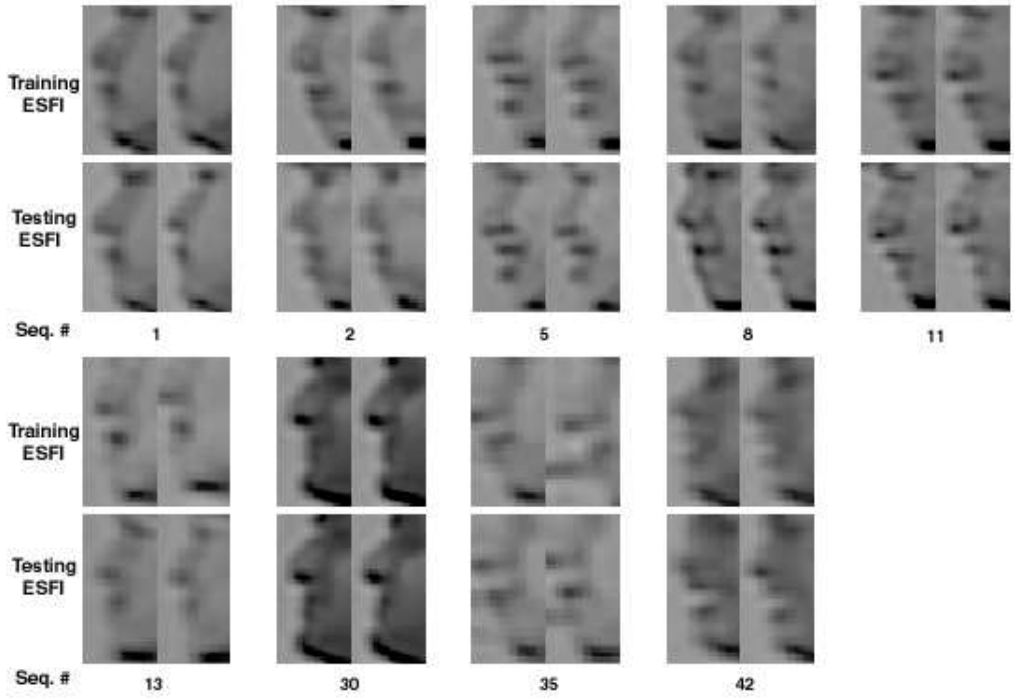
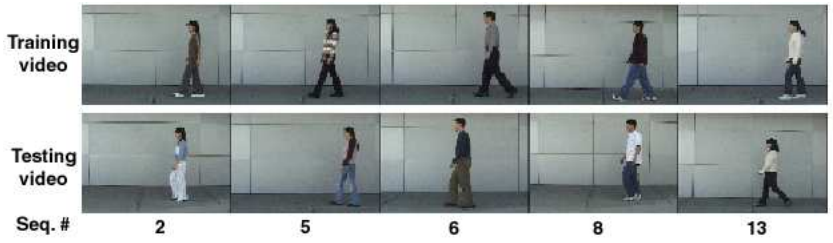
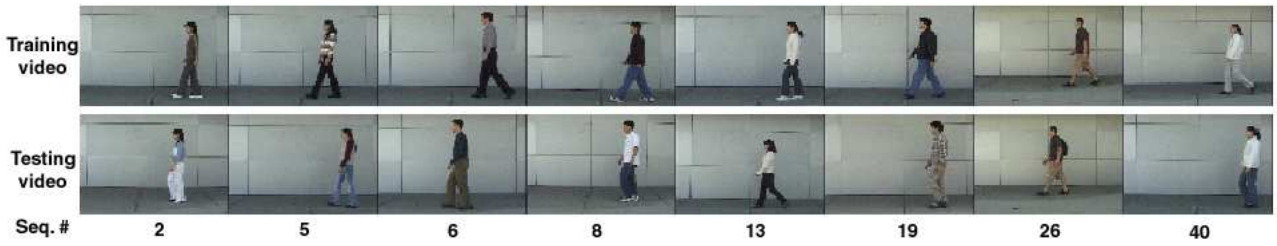


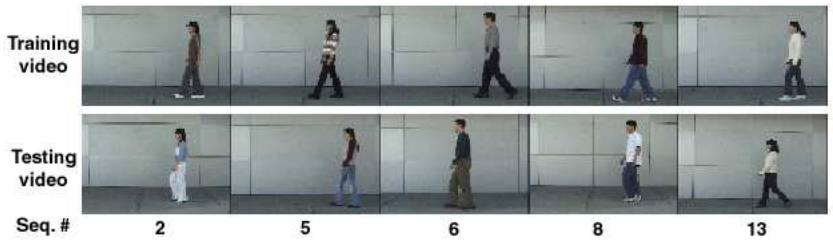
Fig. 14. Experiment 2: ESFIs of people misclassified by the face classifier (see Table V). For each person, 2 ESFIs of the training video sequence and 2 ESFIs of the testing video sequence are shown for comparison.



(a) Errors by Sum rule.



(b) Errors by Product rule.



(c) Errors by Max rule.

Fig. 15. Experiment 2: people misclassified by the integrated classifier based on ESFI and GEI using different fusion rules (see Table VI). For each person, one frame of the training video sequence and one frame of the testing video sequence are shown for comparison.

clothes causes the difference in the shape of the training sequence and the testing sequence for the same person. Also, the lighting condition and the color of clothes cause human body segmentation inaccurate. Figure 13 shows GEIs of people who are misclassified by the gait classifier. Meanwhile, since face is sensitive to noise as well as facial expressions, the different condition in the two video sequences that are taken one month apart, brings face recognition errors. Figure 14 shows ESFIs of people who are misclassified by the face classifier. Note the differences in the training and testing GEIs and ESFIs in Figure 13 and 14. From Table VI, we can see when ESFI and GEI are fused using appropriate fusion methods, the performance improves. Specifically, Sum rule and Max rule based on ESFI and GEI perform the best at the recognition rate of 88.9% (5 errors out of 45 people), and the performance improvement is 6.7% compared with that of the gait classifier. Figure 15 shows people (video sequences) misclassified by integrating ESFI and GEI using different fusion rules. For fusion

based on OSFI and GEI, there is no improvement compared with the individual classifier. These results demonstrate the importance of constructing ESFI. From ESFI, we can extract face features with more discriminating power. Therefore, the performance improvement is still achieved when ESFI instead of OSFI is used for fusion.

3) Experiment 3: The data used in Experiment 3 are the same as the data used in Experiment 2. Experiment 3 studies the effect of using the different number of GEIs and ESFIs in the testing procedure. In the gallery, we still use 2 GEIs and 2 ESFIs for each of the 45 people. While for the probe, we vary the number of GEIs and ESFIs for each person. Table VII shows the performance of fusion by different combination rules when the different number of GEIs and ESFIs is used. Except the performance of fusion using 2 GEIs and 2 ESFIs, which is obtained from Experiment 2, the other performance is the average value on different combination of GEI and ESFI.

TABLE VII

EXPERIMENT 3: FUSED BIOMETRIC PERFORMANCE USING DIFFERENT NUMBER OF GEI AND ESFI.

Fusion Method	1 GEI & 1 ESFI	1 GEI & 2 ESFI	2 GEI & 1 ESFI	2 GEI & 2 ESFI
Sum Rule	82.8%	84.4%	84.4%	88.9%
Product Rule	77.2%	81.1%	82.2%	82.2%
Max Rule	81.1%	80%	84.4%	88.9%

From Table VII, it is clear that if more GEIs and ESFIs are used, i.e., more information in video sequences is used, better performance can be achieved. Meanwhile, this experiment shows that our method to generate the maximum number of synthetic match scores is a reasonable way to use all the available information.

B. Performance Analysis

1) Discussion on Experiments: From Experiments 1 and 2, when ESFI and GEI are used, we can see that Max rule always achieves the best fusion performance, Sum rule has the same fusion performance as the Max rule in Experiment 2, and Product rule does not achieve performance improvement after fusion.

When we compare Experiment 1 and Experiment 2, it can be seen that the recognition rates in Experiment 2 decrease compared with Experiment 1 since 10 out of 45 people change their clothes in the testing sequences. As explained before, gait recognition based on GEI is not only affected by the walking style of a person, but also by the shape of human body. Face is sensitive to noise as well as facial expressions, so the different condition in the training sequence and the testing sequence affects its reliability. All these factors contribute to recognition errors of the individual classifiers. However, the fusion system based on side face and gait overcomes this problem to some extent. In Experiment 2, there are some people who are not correctly recognized by gait, but when side face information is integrated, the recognition rate is improved. It is because the clothes or the walking style of these people are much different between the training and testing video sequences, so the gait classifier can not recognize them correctly. However, the side face of these people does not change so much in the training and testing sequences, and it brings useful information for the fusion system and corrects some errors. Specifically, in Experiment 2, the gait classifier misclassifies 8 people $\{2, 5, 6, 8, 13, 19, 26, 40\}$ and after fusion with ESFI using Sum rule or Max rule, 3 errors $\{19, 26, 40\}$ are corrected. On the other hand, since the face classifier is comparatively sensitive to the variation of facial expressions and noise, it can not get a good recognition rate by itself. When gait information is combined, the better performance is achieved. Our experimental results demonstrate that the fusion system using side face and gait has potential since it integrates cues of side face and gait reasonably, which are two complementary biometrics. Consequently, the fusion system is relatively robust compared with the system using only one individual biometric.

The experimental results in Experiment 1 and 2 clearly demonstrate the importance of constructing ESFI. From ESFI, we can extract face features with more discriminating power. Therefore, better performance is achieved when ESFI instead of OSFI is used for both of the individual classifier and the fused classifier. For example, in Experiment 2, OSFI has bad performance at 64.4%, but ESFI still achieves the recognition rate of 80%. Fusion based on ESFI and GEI achieves the performance improvement of 6.7% (from 82.2% to 88.9%) using the Sum rule and Max rule, while there is no performance improvement by fusion of OSFI and GEI using any combination rule (see Table VI).

Furthermore, from Experiment 3, we can see more information means better performance. This also explains why ESFI always performs better than OSFI since ESFI fuses information from multiple frames.

These results also demonstrate that the match score fusion can not rectify the misclassification achieved by both of the face classifier and the gait classifier. People misclassified by the individual classifiers are likely to be classified correctly after fusion on the condition that there is at least one of the two classifiers that works correctly. For example, in Table V, there are 4 misclassified people $\{2, 5, 8, 13\}$ overlapped between classification using ESFI only and GEI only. There are 8 misclassified people $\{2, 5, 6, 8, 13, 19, 26, 40\}$ overlapped between classification using OSFI only and GEI only. From Table VI, we can see that the set of misclassified people $\{2, 5, 8, 13\}$ are always a subset of the error indices when ESFI and GEI are combined by any fusion rule. Similarly, the set of misclassified people $\{2, 5, 6, 8, 13, 19, 26, 40\}$ are always a subset of the error indices when OSFI and GEI are combined by any fusion rule. It is also the reason that the fusion performance based on OSFI and GEI can never be better than the performance of the gait classifier.

2) Performance Characterization Statistic Q: For the performance improvement by fusion compared with the individual biometric, if the different classifiers misclassify features for the same person, we do not expect as much improvement as in the case where they complement each other [14]. We use a statistic to demonstrate that. There are several methods to assess the interrelationships between the classifiers in a classifier ensemble [15][16]. Given classifiers i and j corresponding to feature vectors f_i and f_j from the same person, respectively, we compute Q statistic:

$$Q_{i,j} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \tag{30}$$

where N^{00} is the number of misclassification by both i and j; N^{11} is the number of correct classification by both i and j; N^{10} and N^{01} are the number of misclassification by i or j, but not by both. It can be easily verified that $-1 \leq Q \leq 1$. The Q value can be considered as a correlation measure between the classifier decisions. The best combination is the one that minimizes the value of Q statistic, which means the smaller the Q value is, the greater the potential for performance improvement by fusion.

TABLE VIII

EXPERIMENT 1: Q STATISTICS.

Fused Templates	N^{11}	N^{00}	N^{01}	N^{10}	Q Statistic
OSFI & GEI	31	1	11	2	0.1698
ESFI & GEI	38	0	4	3	-1

TABLE IX

EXPERIMENT 2: Q STATISTICS.

Fused Templates	N^{11}	N^{00}	N^{01}	N^{10}	Q Statistic
OSFI & GEI	29	8	8	0	1
ESFI & GEI	32	4	5	4	0.7297

Table VIII and IX show the Q values in Experiment 1 and 2. N^{01} is defined as the number of people misclassified by the face classifier but correctly recognized by the gait classifier. N^{10} is defined as the number of people misclassified by the gait classifier but correctly recognized by the face classifier. The Q value based on OSFI and GEI in Experiment 2 is 1, which means the performance improvement by fusion will be zero. The experimental results in Table VI verify it. The Q value based on OSFI and GEI in Experiment 1 is 0.1698, which explains the fact that the fusion performance increases to 95.6% when Product rule is used (see Table IV). When we compare the Q values between fusion of OSFI and GEI, and fusion of ESFI and GEI, the results show that the Q values based on ESFI and GEI are always smaller than the Q values based on OSFI and GEI in both of the experiments. It indicates that the expected performance improvement using ESFI and GEI is higher than using OSFI and GEI. For example, in Experiment 1, the Q value based on fusion of ESFI and GEI is -1 and the Q value based on fusion of OSFI and GEI is 0.1698. The maximum performance increase is 4.5% (from 93.3% to 97.8%) by fusion of ESFI and GEI, while the performance increase by fusion of OSFI and GEI is only 2.3% (from 93.3% to 95.6%). On the other hand, even though the Q value of 0.7297 for fusion performance of ESFI and GEI, are smaller than the Q value of 1 for fusion performance of OSFI and GEI in Experiment 2, it is positive and relatively high. This indicates that many times the gait classifier and the face classifier are both performing correct classification or incorrect classification for

the same person. In spite of this, our video based fusion method using ESFI and GEI always achieves better performance than either of the individual classifier when the appropriate fusion strategy is used.

IV. Conclusions

This paper proposes an innovative video based fusion system, which aims at recognizing non-cooperating individuals at a distance in a single camera scenario. Information from two biometric sources, side face and gait, is combined using different fusion methods. Side face includes the entire side views of eye, nose and mouth, possessing both shape information and intensity information. Therefore, it has more discriminating power for recognition than face profile. To overcome the problem of limited resolution of side face at a distance in video, we use Enhanced Side Face Image (ESFI), a higher resolution image constructed from multiple video frames instead of OSFI directly obtained from a single video frame, as the face template for an individual. ESFI serves as a better face template than OSFI since it generates more discriminating face features. Synthetic match scores are generated for fusion based on the characteristics of face and gait. The experimental results show that the integration of information from side face and gait is effective for individual recognition in video. The performance improvement is always archived when appropriate fusion rules, such as the Max rule and the Sum rule, are used to integrate information from ESFI and GEI. Consequently, our fusion system is relatively robust compared with the system using only one biometric in the same scenario.

However, our system has some limitations: (a) Gait recognition based on GEI is affected by the shape of human body to some extent; (b) The side face contains less information compared with the frontal face and it is sensitive to noise as well as facial expressions; (c) The system has been tested on limited video sequences. In spite of these limitations, we demonstrate that the integration of face and gait can achieve better recognition performance at a distance in video. Although our database is not very big, but it is of reasonable size (45 people with 100 video sequences) and shows how the proposed ideas work. In the future, we will collect more data to evaluate the performance of our system. We will also focus on problems that are not addressed in this paper. We will use multiple cameras to capture different views of a person. To get face images with high quality, we will track

the whole human body first and then zoom in to get better face images. We will speed up the process of ESFI and GEI construction so that our system can operate in real time.

REFERENCES

[1] A. Kale, A. Roychowdhury, and R. Chellappa, "Fusion of gait and face for human identification," in *Proc. Acoustics, Speech, and Signal Processing 2004*, vol. 5, 2004, pp. 901–904.

[2] G. Shakhnarovich and T. Darrell, "On probabilistic combination of face and gait cues for identification," in *Proc. Automatic Face and Gesture Recognition 2002*, vol. 5, 2002, pp. 169–174.

[3] G. Shakhnarovich, L. Lee, and T. Darrell, "Integrated face and gait recognition from multiple views," in *Proc. Computer Vision and Pattern Recognition 2001*, vol. 1, 2001, pp. 439–446.

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

[5] J. Han and B. Bhanu, "Performance prediction for individual recognition by gait," *Pattern Reconition Letters*, vol. 26, no. 5, pp. 615–624, 2005.

[6] P. A. Hewitt and D. Dobberfuhl, "The science and art of proportionality," *Science Scope*, pp. 30–31, 2004.

[7] S. Periaswamy and H. Farid, "Elastic registration in the presence of intensity variations," *IEEE Transactions on Medical Imaging*, vol. 22, no. 7, pp. 865–874, 2003.

[8] B. Horn, *Robot Vision*. MIT Press, Cambridge, MA, 1986.

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

[10] B. Bhanu and X. Zhou, "Face recognition from face profile using dynamic time warping," in *17th International Conference on Pattern Recogntion*, vol. 4, 2004, pp. 499–502.

[11] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. PAMI*, vol. 28, no. 2, pp. 316–322, 2006.

[12] J. J. Little and J. E. Boyd, "Recognizing people by their gait: the shape of motion," *Videre: Journal of Computer Vision Research*, vol. 1, no. 2, pp. 1–32, 1998.

[13] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[14] T. Kinnune, V. Hautamaki, and P. Franti, "Fusion of spectral feature sets for accurate speaker identification," in *Proc. 9th International Conference Speech and Computer (SPECOM'2004)*, September 2004, pp. 361–365.

[15] C. A. Shipp and L. I. Kuncheva, "Relationships between combination methods and measures of diversity in combining classifiers," *Information Fusion*, vol. 3, pp. 135–148, 2002.

[16] G. Brown, J. Wyatt, R. Harris, and X. Yao, "Diversity creation methods: a survey and categorisation," *Information Fusion*, vol. 6, no. 1, pp. 5–20, 2005.