10.6 Conclusion

We have presented a general algorithmic solution for 3D object recognition, and how to adapt it for 3D face and 3D ear recognition. By utilizing a deformable model we map the 3D geometry information onto a 2D regular grid, thus combining the rich information of the 3D data with the computational efficiency of 2D data. A multistage fully automatic alignment algorithm and the advanced wavelet analysis resulted in state-of-the-art performance on the publicly available FRGC v.2 face database. We have also presented encouraging results on our own 3D ear database and on a subset of the publicly available UND ear database.

The current trend in biometrics is to achieve higher accuracy and robustness by using multiple biometric modalities. It is becoming increasingly accepted that no biometric modality can provide 100% verification rate at very low FARs in large databases. Any hope for achieving such an accuracy figure can only result from the fusion of multiple modalities. We have observed the same phenomenon in 3D face recognition at a smaller scale: our verification rates increased significantly by fusing position data and normal data as well as fusing Haar and Pyramid wavelet transforms of the above two types of data. We expect that the fusion of 3D face and 3D ear biometric data from same individual will provide good fusion results. To test this hypothesis a large dataset of 3D face and 3D ear data needs to be acquired. An additional advantage of such a fusion is that the cost of the 3D capture device can be amortized over these two modalities.

11 Human Recognition at a Distance in Video by Integrating Face Profile and Gait

Xiaoli Zhou, Bir Bhanu, and Ju Han

11.1 Introduction

It has been found to be difficult to recognize a person from arbitrary views in reality, especially when one is walking at a distance in real-world outdoor conditions. For optimal performance, the 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 3D models. However, the problem of building reliable 3D models for nonrigid 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 3D models have achieved some progress. In [296], Kale et al. present a gait recognition algorithm and a face recognition algorithm based on sequential importance sampling. The fusion of frontal face and gait cues is performed in the single camera scenario. In [297, 298], Shtaknarovich et al. compute an image-based visual hull from a set of views of four monocular cameras. It is then used to render virtual canonical views for tracking and recognition. The gait recognition scheme is based on silhouette extent analysis. Eigenfaces are used for recognizing frontal face rendered by the visual hull. They discuss the issues of crossmodal correlation and score transformations for different modalities and present the fusion of face and gait.

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 are easily satisfied by an individual classifier for face or gait, but pose some challenges when one attempts to integrate face and gait biometrics in real-world applications. In Kale's and Shakhnarovich's fusion systems [296-298], both use the side view of gait and the frontal view of face. In Kale's work [296], the subjects are 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 gallery consists of static faces for the corresponding subjects. Therefore, they perform still-to-video face recognition. In Shakhnarovich's work [297, 298], four cameras must be used to get both the canonical view of gait and the frontal view of face simultaneously.
In this chapter, an innovative system is proposed, aiming at recognizing non-cooperating individuals at a distance in a single camera scenario. Information from two biometric sources, face profile and gait, is combined. We use face profile instead of frontal face in the system since a side view of face is more likely to be seen than a frontal view of a face when one exposes the best side view of gait to the camera. It is very natural to integrate information of the side face view and the side gait view. However, it is difficult to get reliable information of a face profile directly from a low-resolution video frame for recognition tasks because of limited resolution. To overcome this problem, we use resolution enhancement algorithms for face profile analysis. We first reconstruct a high-resolution face profile image from multiple adjacent low-resolution video frames. The high-resolution face profile image fuses both the spatial and temporal information present in a video sequence. The approach 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 for face profile [299]. Then, we extract face profile features from the high-resolution face profile images. Finally, in a dynamic time warping (DTW) method [300] is used to match face profiles based on absolute values of curvature. For gait, we use gait energy image (GEI), a spatiotemporal compact representation, to characterize human walking properties [301]. Recognition is carried out based on the direct GEI matching. Face profile cues and gait cues are integrated by three schemes. The first two are Sum rule and Product rule [302]. The last one is an indexing-verification scheme, which consolidates the accept/reject decisions of multiple classifiers [303].

This chapter is organized as follows. Section 11.2 presents the overall technical approach. It explains the construction of a high-resolution face profile image and describes the generation of GEI. It presents the details of face profile recognition and gait recognition. It provides a description of the fusion of face profile and gait, and the classification methods. In Sect. 11.3, a number of dynamic video sequences are tested. Experimental results are compared and discussed. Finally, Sect. 11.4 concludes this chapter.

11.2 Technical Approach

The overall technical approach is shown in Fig. 11.1. A simple background subtraction method [304] is used for human body segmentation from video data. For each video sequence in the gallery, we construct a high-resolution face profile image from low-resolution face profile images, and a GEI from the binary silhouette image sequences. Then, we extract face profile features from each high-resolution profile image to form face feature gallery. During the testing procedure, each testing video is processed to generate both the high-resolution face profile image and the GEI. The face profile features are extracted from the high-resolution face profile image and compared with face profile features in the gallery using DTW. The GEI is directly compared with the GEI templates in the gallery. Finally, different fusion strategies are used to combine the results of the face profile classifier and the gait classifier to improve recognition performance.

11.2.1 High-Resolution Image Construction for Face Profile

Multiframe resolution enhancement seeks to construct a single high-resolution image from multiple low-resolution images. These 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. The idea of super-resolution was first introduced for multiframe image restoration of band-limited signals in 1984 [305]. In the last two decades, different mathematical approaches have been developed. All of them seek to address the question of how to combine redundant image information present in multiple images.

In this chapter, the original low-resolution face profile images are first localized and extracted from the segmented human body obtained from multiple video frames. A human body is divided into two parts according to the proportion of its parts [306]: 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 obtain the original low-resolution face profile images by cutting the upper 10% of the segmented human body. Before multiple low-resolution face images are fused to construct a high-resolution face image using the resolution enhancement method, they are aligned by affine transformation and motion estimates are computed to determine pixel displacements between them. Then, an iterative method [307] is used to construct a high-resolution face profile image from aligned low-resolution face profile images.
The Imaging Model

The imaging process, yielding the observed face profile image sequence $f_k$, is modeled by [307]

$$f_k(m,n) = \sigma_k(h(T_k(F(x,y))) + \eta_k(x,y))$$  \hspace{1cm} (11.1)

where

1. $f_k$ is the sensed image of the tracked face profile in the kth frame.
2. $F$ is a high-resolution image of the tracked face profile in a desired reconstructed view. Finding $F$ is the objective of the super-resolution algorithm.
3. $T_k$ is the 2D geometric transformation from $F$ to $f_k$, determined by the 2D motion parameters of the tracked face profile in the image plane. $T_k$ is assumed to be invertible and does not include the decrease in the sampling rate between $F$ and $f_k$.
4. $h$ 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.
5. $\eta_k$ is an additive noise term.
6. $\sigma_k$ is a downsampling 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 $F$, which approximates $F$ as accurately as possible, and surpasses the visual quality of the observed images in $\{f_k\}$.

The Super-Resolution Algorithm

The algorithm for creating higher resolution images is iterative. Starting with an initial guess $F^{(0)}$ for the high-resolution face profile image, the imaging process is simulated to obtain a set of low-resolution face profile 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 face profile image, then the simulated images $\{f_k^{(0)}\}_{k=1}^K$ should be identical to the observed low-resolution face profile images $\{f_k\}_{k=1}^K$. The difference images $f_k^{(0)} - f_k^{(0)}\_{k=1}$ 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 face profile image $F^{(1)}$. This process is repeated iteratively to minimize the error function

$$e^{(n)} = \sqrt{\frac{1}{K} \sum_{k=1}^K ||f_k - f_k^{(n)}||^2}$$  \hspace{1cm} (11.2)

11.2 Recognition by Integrating Face Profile and Gait

The imaging process of $f_k$ at the n-th iteration is simulated by

$$f_k^{(n)} = (T_k(F^{(n)}) * h) \downarrow s$$  \hspace{1cm} (11.3)

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

$$F^{(n+1)} = F^{(n)} + \frac{1}{K} \sum_{k=1}^K T_k^{-1}(f_k - f_k^{(n)}) \uparrow s * p$$  \hspace{1cm} (11.4)

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

In our system, we reconstruct a high-resolution face profile image from six adjacent video frames. We assume that six low-resolution face profile images have been localized and extracted from adjacent video frames. We then align these six low-resolution face profile images using affine transformation. Affine transformation works for in-plane, not out of plane rotations of the human face. The quality of the reconstructed image depends on how well the six profile images are registered. Finally, we apply the super-resolution algorithm given above to construct a high-resolution face profile image from the six aligned low-resolution face profile images. The resolution of the original low-resolution face profile images is $70 \times 70$ and the resolution of the reconstructed high-resolution face profile image is $140 \times 140$. Figure 11.2 shows the six low-resolution face profile images from six adjacent frames:

(a) (b) (c) (d) (e) (f)

Fig 11.2. Six low-resolution face profile images resized by using bilinear interpolation (a-f)
easy to analyze and more foolproof. Within the last decade, several algorithms have been proposed for automatic person identification using face profile images. Most of these algorithms depend on the correct detection of all fiducial points and the determination of relationships among these fiducial points.

Harmon and Hunt [308] use manually entered profile traces from photographs of 256 male faces. They locate eight independent fiducials on the profiles and obtain the ninth fiducial by rotating a point from the chin about the prominence until it intersects the profile above the prominence. Later, Harmon et al. [309] increase the number of fiducials from nine to eleven, and achieve 96% recognition accuracy for 112 subjects, using a 17-dimensional feature vector. The most significant problem with tangency-based techniques is that there is not a line that is tangent to the prominence and chin for profiles with protruding lips [310]. Campos et al. [311] analyze the profile of the face using scale-space techniques to extract eight fiducials. This technique assumes that there will be nine zero-crossings on the profile, and this assumption could be invalidated by facial hair particularly moustaches and the hairline on the forehead. Darwish et al. [312] extract nine fiducials based on the observation that the curvature of the profile alternates between convex and concave, with the point of maximal absolute curvature in each segment corresponding to a fiducial. Akimoto et al. [313] use a template matching approach to find the position of the same five fiducials used by Galton [314]. The template consisting of approximately 50 line segments is used to represent a generic face profile.

In reality, some profiles are too difficult for all fiducials to be reliably extracted, so in these cases a feature vector approach based on the same fiducial points of different face profiles will fail. In this chapter, we use a curvature-based matching approach [300] for recognition, which does not focus on the extraction of all the fiducial points and the determination of relationship among these fiducial points [309, 311, 312]. We use the relationship of some fiducial points for their extraction, but not for an individual recognition. The Gaussian scale-space filter is first used to smooth the profile extracted from the high-resolution face profile image and then the curvature of the filtered profile is computed. Using the curvature values, the fiducial points, including the nasion and the orbit, can be reliably extracted using a fast and simple method after prominence is determined. Finally, a DTW method is applied to compare the face profile portion from nasion to orbit based on the curvature values.

Face Profile Representation

We apply a canny edge detector to the high-resolution face profile image. After edge linking and thinning, the profile of a face is extracted as the lefmost points different from background, which contain fiducial points like nasion, prominence, and orbit. The outline of a profile is treated as a 1D function, consisting of a set of points \( T = (x, y) \), where \( x \) is a row index and \( y \) is a column index of a pixel. The Gaussian scale-space filter is applied to the face profile to eliminate the spatial quantization noise introduced during the digitization process, as well as other types of high frequency noise. The convolution between Gaussian kernel \( g(x, \sigma) \) and signal

**Fig. 11.4.** The reconstructed high-resolution face profile and its edge image

video frames. For comparison, we resize the six low-resolution face profile images by using bilinear interpolation. Figure 11.3 shows the corresponding edge images of six low-resolution face profiles. Figure 11.4 shows the reconstructed high-resolution face profile image and its edge image. From these figures, we can see that the reconstructed high-resolution image is better than any of the six low-resolution images. It is clearly shown in the edge images that the edges of the high-resolution image are much smoother and more accurate than that of the low-resolution images. Using the reconstructed high-resolution image, we can extract better features for face profile matching.

11.2.2 Face Profile Recognition

Face profile is an important aspect for the recognition of faces, which provides a complementary structure of the face that is not seen in the frontal view. Though it inherently contains less discriminating power than frontal images, it is relatively
The function $f(x)$ depends both on $x$, the signal's independent variable, and on $\sigma$, the Gaussian's standard deviation. It is given by

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

(11.5)

where $\ast$ denotes convolution with respect to $x$. The bigger the $\sigma$, 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 $T_{e}(u) = (X(u, \sigma), Y(u, \sigma))$, where $X(u, \sigma) = x(u) \ast g(u, \sigma)$ and $Y(u, \sigma) = y(u) \ast g(u, \sigma)$.

Curvature $\kappa$ on $T_e$ is computed as

$$\kappa(u, \sigma) = \frac{X_u(u, \sigma)Y_{uu}(u, \sigma) - X_{uu}(u, \sigma)Y_u(u, \sigma)}{(X_u(u, \sigma)^2 + Y_u(u, \sigma)^2)^{3/2}}$$

(11.6)

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

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

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

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

Since the profiles include the hair and some other parts that are not reliable for matching, we extract a portion of profile starting from nasion to thrust for effective matching. It is done by finding the fiducial points on the face profile. 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 too many parameters and does not require any thresholds. It is also fast and simple.

We define pronasale as the leftmost point above thrust in the middle part of the profile and nasion as the first point that has local maximum of the absolute values above pronasale. The method of extracting the nasion and thrust points is described as follows:

1. Compute the curvature of a profile at an initial scale, find local maxima of the absolute values as corner candidates, and track them down to lower scales.
2. Regard the rightmost point in the candidate set as the thrust.
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.
4. Assume that there are no candidate points between pronasale and nasion and identify the first candidate point above the pronasale as nasion.

Figure 11.5 shows the extracted face profile and the absolute values of curvature. It is clear that the locations of the fiducial points, including nasion, pronasale, and thrust, have local maxima of the absolute values. Figure 11.6 shows the absolute values of curvature on face profiles belonging to four different people. We can see that different face profiles have different patterns of curvature. Therefore, we can use the absolute values of curvature as the feature to represent a face profile. Curvature features have some advantages in that they are invariant to rotation, translation, and uniform scaling.

Face Profile Matching Using Dynamic Time Warping

We use the Dynamic Time Warping (DTW) as the matching method to compute the similarity of two face profiles based on the absolute values of curvature, which are used to represent the shapes of face profiles. The DTW is an algorithm to calculate the optimal score and to find the optimal alignment between two strings. This method is a much more robust distance measure for time series than Euclidean distance, allowing similar shapes to match even if they are out of phase in the time axis [315]. We use the Needleman–Wunsch [316] global alignment algorithm to find the optimum alignment of two sequences when considering their entire length. For two sequences $s(1 \ldots n)$ and $t(1 \ldots m)$, we compute $D(i, j)$ for entire sequences, where $i$ ranges from 1 to $m$ and $j$ ranges from 1 to $n$. $D(i, j)$ is defined as
similarity matrix in Fig. 11.7, we can see a light stripe (high similarity values) approximately down the leading diagonal. From the dynamic programming matrix in Fig. 11.7, we can see the lowest-cost path between the opposite corners visibly follows the light stripe, which overlay the path on the similarity matrix. The least cost is the value in the bottom-right corner of the dynamic programming matrix. This is the value we would compare between different templates when we are doing classification. The unknown person is classified to the class which gets the least cost out of all the costs corresponding to all the classes.

11.2.3 Gait Recognition

In recent years, various techniques have been proposed for human recognition by gait. These techniques can be divided as model-based and model-free approaches. In this chapter, we focus on a model-free approach that does not recover a structural model of human motion. In the following, we provide related work on gait recognition.

Little and Boyd [317] describe the shape of the human motion with scale-independent features from moments of the dense optical flow, and recognize individuals by phase vectors estimated from the feature sequences. Sundaram et al. [318] propose a hidden Markov models (HMMs)-based framework for individual recognition by gait. Huang et al. [319] extend the template matching method to gait recognition by combining transformation based on canonical analysis and eigenspace transformation for feature selection. Sukkar et al. [320] directly measure the similarity between the testing and training sequences by computing the correlation of corresponding time-normalized frame pairs. Collins et al. [321] first extract key frames from a sequence and then compute the similarity between two sequences using the normalized correlation.

While some gait recognition approaches [319] extract features from the correlation of all the frames in a walking sequence without considering their order, other approaches extract features from each frame and compose a feature sequence for the human walking sequence [317, 320, 321]. During the recognition procedure, these
approaches either match the statistics collected from the feature sequence, or match
the features between the corresponding pairs of frames in two sequences that are
time-normalized with respect to their cycle lengths. The fundamental assumptions
made here are (1) the order of poses in human walking cycles is the same, i.e.,
limbs move forward and backward in a similar way among normal people, and (2)
differences exist in the phase of poses in a walking cycle, the extend of limbs, and
the shape of the torso, etc. Under these assumptions, it is possible to represent the
spatiotemporal information in a single 2D gait template, called gait energy image
(discussed below), instead of an ordered image sequence.

Gait Frequency and Phase Estimation

Regular human walking can be considered as cyclic motion where human motion
repeats at a stable frequency. Therefore, it is possible to divide the whole gait
sequence into cycles and study them separately. We assume that silhouette extract-
ion has been performed on original human walking sequences, and 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 part size from
each frame indicates the gait frequency and phase information. We estimate the gait
frequency and phase by maximum entropy spectrum estimation [317] from the
time series signal.

Gait Representation

Given the preprocessed binary gait silhouette image $I_t(x,y)$ at time $t$ in a sequence,
the gray-level gait energy image (GEI) is defined as follows [301]:

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

(11.9)

where $N$ is the number of frames in the complete cycle(s) of a silhouette sequence,
$t$ is the frame number of the sequence (moment of time), and $x$ and $y$ are values
in the 2D image coordinate. Figure 11.8 shows the sample silhouette images in a
gait cycle from two 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 accumula-
tive energy image of human walking in the complete cycle(s), and (c) a pixel with
higher intensity value in GEI means that human walking occurs more frequently at
this position (i.e., with higher energy).
\[ \hat{S}_i = \frac{\exp(-S_i)}{\sum_{l=1}^{c} \exp(-S_l)} \quad i = 1, 2, \ldots, c \]  
(11.11)

After normalization, the match scores of face and gait from the same class are fused using different fusion methods. Let \( \hat{S}_i^f \) and \( \hat{S}_i^g \) be the normalized face match scores and the normalized gait match scores, respectively. The unknown person is classified to class \( k \) if

\[ R(\hat{S}_k^f, \hat{S}_k^g) = \max R(\hat{S}_i^f, \hat{S}_i^g) \]  
(11.12)

where \( R(\cdot, \cdot) \) means a fusion method. Sum and Product rules [302] are used in our experiments. Distances representing dissimilarity become match scores representing similarity by using (11.11), so the unknown person is classified to the class which gets the largest integrated match score out of all the integrated match scores corresponding to all the classes.

The last one is an indexing-verification scheme. In a biometric fusion system, a less accurate, but fast and simple classifier can pass on a smaller set of candidates to a more accurate, but time-consuming and complicated classifier. In our system, the face profile classifier works as a filter to pass on a smaller set of candidates to the next stage of gait classifier. Then, the gait classifier compares similarity among these candidates based on GEIs. The result of the gait classifier is the result of the fusion system.

11.3 Experimental Results

11.3.1 Data

The data are obtained by a Sony DCR-VX1000 digital video camera recorder. We collect 28 video sequences of 14 people walking in the outdoor conditions and exposing a side view to the camera. The camera operates at about 30 frames s\(^{-1}\). The resolution of each frame is 720 × 480. The distance between people and the video camera is about 10 ft. Each of people has two sequences, one for training and the other one for testing. Each sequence includes one person. Figure 11.9 shows some video frames of four people.

11.3.2 Experiments

From each sequence, we construct one high-resolution face profile image from six low-resolution face profile images that are extracted from six adjacent video frames, and one GEI from a complete walking cycle that includes about 20 video frames. Since there are two sequences for each of 14 people, we have 14 high-resolution face profile images and 14 GEIs in the gallery, and another 14 high-resolution face profile images and 14 GEIs in the probe. The resolution of low-resolution face profile images is 70 × 70 and the resolution of reconstructed high-resolution face profile images is 140 × 140. The resolution of each GEI is 128 × 88.

Recognition metric is used to evaluate the performance of our method, the quality of extracted features, and their impact on identification. It is defined as the ratio of the number of the correctly recognized people to the number of all the people. The results for our database are shown in Table 11.1. We can see that 64.3\% people are correctly recognized (5 errors out of 14 persons) by face profile and 85.7\% people are correctly recognized by gait (2 errors out of 14 persons), respectively. For the fusion schemes, the best performance is achieved by the Sum rule at 100% accuracy. The Product rule and the indexing-verification scheme obtain the same recognition rate at 92.9%. When we use the indexing-verification scheme, we choose the first three matching results of the face profile classifier as candidates. Then, the gait classifier measures the similarity between the corresponding GEI of the testing people and the corresponding GEI of the training people in the candidate list. The unknown person is finally classified as the most similar class among the candidates.

From Table 11.1, we can see that there are two people who are not correctly recognized by gait, but when the face profile classifier is integrated, the recognition rate is improved. It is because 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. Environmental and clothing changes cause the difference in the shape of the training sequence and
the testing sequence for the same person. However, the face profiles of these two people do not change so much in the training and the testing sequences. It shows that face profile is a useful cue for the fusion system. Figure 11.10 shows the corresponding GEIs of two people who are misclassified by the gait classifier. Figure 11.11 shows the corresponding face profiles of two people who are misclassified by the gait classifier. Note the difference in the training and testing GEIs in Fig. 11.10 and

Fig. 11.10. GEIs of two people misclassified by the gait classifier. For each person, the training GEl and the testing GEl are shown for comparison.

Fig. 11.11. Face profile of two people misclassified by the gait classifier. For each person, the training profile and the testing profile are shown for comparison.

the similarity in the training and testing face profiles in Fig. 11.11. Since the face profile classifier is comparatively sensitive to the variation of facial expression and noise, the face profile classifier cannot get a good recognition rate by itself. When the gait classifier is combined with the face profile classifier, the better performance is achieved.

From the experiments, we can see that the fusion system using face profile and gait is promising. The fusion system has better performance than either of the individual classifier. It shows that our fusion system is relatively robust. Although the experiments are only done on a small database, our system has the potential since it integrates cues of face profile and cues of gait reasonably, which are independent biometrics.

11.4 Conclusions

This chapter introduces a video-based system combining face profile and gait for human recognition in a single-camera scenario. For optimal face profile recognition, we extract face profile features from a high-resolution face profile image constructed from multiple video frames instead of a low-resolution face profile image directly obtained from a single video frame. For gait recognition, we use GEl, a spatiotemporal compact representation to characterize human walking properties. Several schemes are considered for fusion of face profile and gait. The experimental results show that the integration of information from face profile and gait is effective for individual recognition in video. The performance improvement is archived when appropriate fusion rules are used. The idea of constructing the high-resolution face profile image from multiple video frames and generating the GEl is promising for human recognition in video.

Several issues that concern real-world applications require further research in the future. These include the extraction of accurate face profile from video frames in a crowded surveillance application, extraction of reliable silhouettes of moving people in the presence of environmental and clothing changes, and real-time operation of the fusion system. Moreover, for face profile recognition, the outer contour of the side face is sensitive to local distortion and noise. In our recent work [322, 323], we have used the side face, which includes entire side views of eye, nose, and mouth (discarding facial hair). Since it possesses both the shape and the intensity information, it is found to have more discriminating power for recognition than a face profile.
Face Biometrics for Personal Identification

Multi-Sensory Multi-Modal Systems