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

Xiaoli Zhou, Bir Bhanu, and Ju Han

Center for Research in Intelligent Systems
University of California, Riverside CA 92521, USA
{xzhou,bhanu,jhan} @vislab.ucr.edu

Abstract. Human recognition from arbitrary views is an important task for many applications, such as visual surveillance, covert security and access control. It has been found to be very difficult in reality, especially when a person is walking at a distance in read-world outdoor conditions. For optimal performance, the system should use as much information as possible from the observations. In this paper, we propose an innovative system, which combines cues of face profile and gait silhouette from the single camera video sequences. For optimal face profile recognition, we first reconstruct a high-resolution face profile image from several adjacent low-resolution video frames. Then we use a curvature-based matching method for recognition. For gait, we use Gait Energy Image (GEI) to characterize human walking properties. Recognition is carried out based on the direct GEI matching. Several schemes are considered for fusion of face profile and gait. A number of dynamic video sequences are tested to evaluate the performance of our system. Experiment results are compared and discussed.

1 Introduction

It has been found to be very 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 low-resolution video sequences, is a practical approach to accomplish the task of human recognition.

The most general solution to analyze face and gait information from arbitrary views is to estimate 3-D models. However, the problem of building reliable 3-D models for articulating objects like the human body remains a hard problem. In recent years, the way to perform integrated face and gait recognition without resorting to 3-D models has made some progress. In [1], Kale et al. present a view invariant gait recognition algorithm and a face recognition algorithm based on sequential importance sampling. The fusion of frontal face and gait cues is in the single camera scenario. In [2], 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. A gait recognition scheme is based on silhouette extent analysis. Eigenfaces are used for recognizing frontal face rendered by the visual hull. In a later work [3], Shakhnarovich et al. discuss the issues of cross-modal correlation and score transformations for different modalities and present the probabilistic settings for the cross-modal fusion.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 533–543, 2005.

Most 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 analyzed. The requirement of different views is easily satisfied by an individual classifier, while it brings some difficulties to the fusion system. In Kale's and Shakhnarovich's fusion system, both of them use the side view of gait and the frontal view of face. So in Kale's work [1], only the final segment of the NIST database can present a nearly frontal view of face, while in Shakhnarovich's work [2][3], multiple cameras must be used to get both the side view of gait and the frontal view of face simultaneously.

In this paper, an innovative system is proposed, which combines cues of face profile and gait silhouette from the single camera video sequences. We use face profile instead of frontal face in the system since a side view of face is more probable to get than a frontal view of a face when one exposes the best side view of gait to the cameral. It is very natural to integrate information of the side face view and the side gait view. However, it is hard to get enough information of a face profile directly from a low-resolution video frame for recognition tasks. To overcome this limitation, we use super-resolution algorithms for face profile analysis. We first reconstruct a high-resolution face profile image from several adjacent low-resolution video frames. 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. Then we extract good features from the high-resolution face profile images. Finally, a curvature-based matching method is applied [4]. For gait, we use Gait Energy Image (GEI) to characterize human walking properties [5]. Recognition is carried out based on the direct GEI matching.

Face profile cues and gait cues are considered being integrated by several schemes. The first two are SUM rule and PRODUCT rule [6]. We assume features of face profile and features of gait we use statistically independent, so matching scores reported by the individual classifier can be combined based on Bayesian Theory. The last one is an indexing-verification scheme, which consolidates the accept/reject decisions of multiple classifiers [7]. The overall technical approach is shown in Fig. 1.

2 Technical Approach

2.1 High-Resolution Image Construction for Face Profile

Multiframe resolution enhancement, or super-resolution, seeks to construct a single high-resolution image from several 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 in 1984 by Tsai and Huang [8] for multiframe image restoration of band-limited signals. In the last two decades, different mathematical approaches have been developed. All of them seek to address the question of how to combine irredundant image information in multiple frames. A good overview of existing algorithms is given by Borman and Stevenson [9] and Park et al. [10]. In this paper, we use an iterative method proposed by Irani and Peleg [11][12].

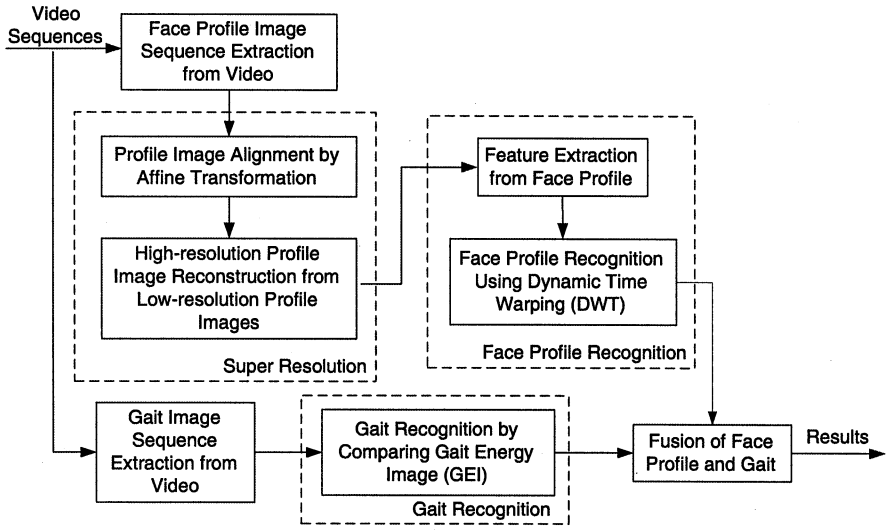

Fig. 1. Technical approach for integrating face profile and gait in video

The Imaging Model. The imaging process, yielding the observed image sequence g_k, is modeled by:

$$g_k(m,n) = \sigma_k(h(T_k(f(x,y))) + \eta_k(x,y)) \tag{1}$$

where

1. g_k is the sensed image of the tracked object in the k_{th} frame.
2. f is a high resolution image of the tracked object in a desired reconstruction view. Finding f is the objective of the super-resolution algorithm.
3. T_k is the 2-D geometric transformation form f to g_k, determined by the computed 2-D motion parameters of the tracked object in the image plane (not including the decrease in sampling rate between f and g_k). T_k is assumed to be invertible.
4. h is a blurring operator, determined by the Point Spread Function of the sensor (PSF). When lacking knowledge of the sensor's properties, it is assumed to be a Gaussian.
5. η_k is an additive noise term.
6. σ_k is a downsampling operator which digitizes and decimates the image into pixels and quantizes the resulting pixels values.

The receptive field (in f) of a detector whose output is the pixel $g_k(m,n)$ 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 $T_k^{-1}((m,n))$. An attempt is made 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 $\{g_k\}$.

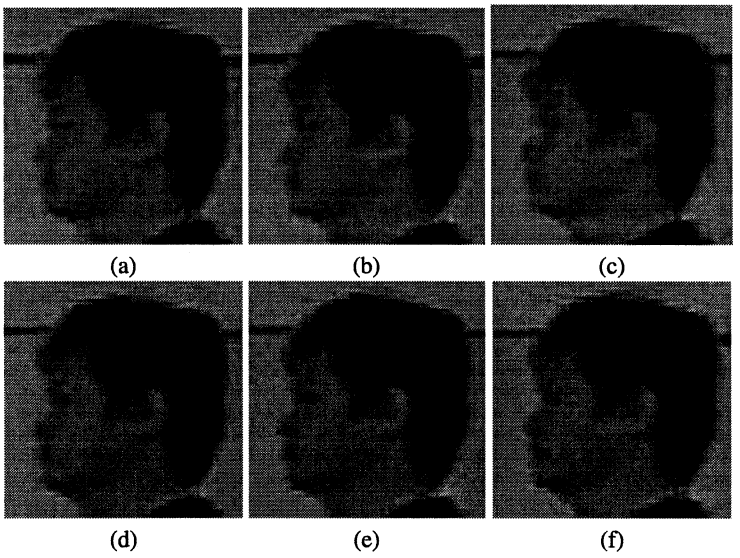

Fig. 2. The six low-resolution face profile images resized by using bilinear interpolation (a-f)

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

$$e^{(n)} = \sqrt{\frac{1}{K}\sum_{k=1}^{K} \|g_k - g_k^{(0)}\|_2^2} \tag{2}$$

The imaging process of g_k at the n_{th}iteration is simulated by:

$$g_k^{(n)} = (T_k(f^{(n)}) * h) \downarrow s \tag{3}$$

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:

$$f^{(n+1)} = f^{(n)} + \frac{1}{K}\sum_{k=1}^{K} T_k^{-1}(((g_k - g_k^{(n)}) \uparrow s) * p) \tag{4}$$

where K is the number of low resolution images. $\uparrow s$ is an upsampling operator by a factor s, and p is a "backprojection" kernel, determined by h and T_k as explained below.

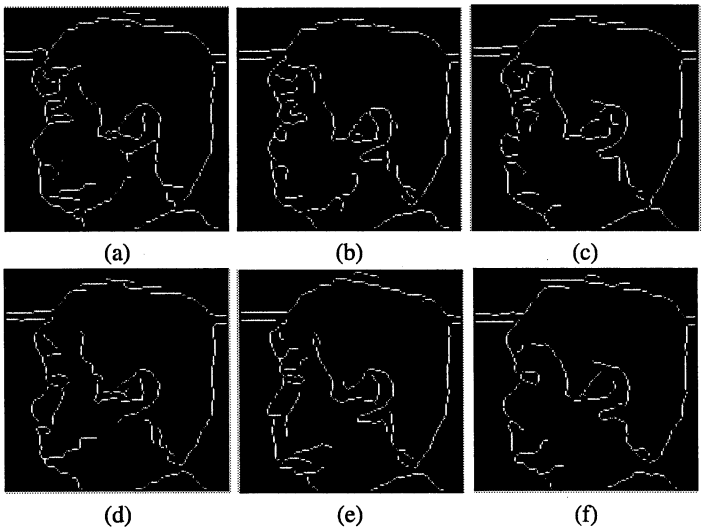

Fig. 3. The edge images of six low-resolution face profiles

Fig. 4. The reconstructed high-resolution face profile and its edge image

The averaging process reduces additive noise. The algorithm is numerically similar to common iterative methods for solving sets of linear equations, and therefore has similar properties, such as rapid convergence.

In our system, we reconstruct a high-resolution face profile image from six adjacent video frames. It 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. We assume that six low-resolution face profile images have been localized and extracted from six adjacent video frames. We then align these six low-resolution face profile images using affine transformation. Finally, we apply the super resolution algorithm 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 70x70 and the resolution of the reconstructed high-resolution face profile image is 140x140. Figure 2 shows the six low-resolution face profile images from six adjacent video frames. For comparison, we resize the six low-resolution face profile images by using bilinear interpolation. Figure 3 shows the corresponding edge images of six low-resolution face profiles. Figure 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 much 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 reliable than that of the low-resolution images. This explains why we need to apply super resolution algorithm to our problem. Using the reconstructed high-resolution image, we can extract good features for face profile matching.

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. For face profile recognition, we use a curvature-based matching approach [4], which does not focus on all fiducial point extraction and the determination of relationship among these fiducial points like most of current algorithms do, but attempt to use as much information as a profile possesses. The scale space filtering is used to smooth the profile and then the curvature of the filtered profile is computed. Using the curvature value, the fiducial points, including the nasion and throat can be reliably extracted using a fast and simple method after pronasale is decided. Then a dynamic time warping method is applied to compare the face profile portion from nasion to throat based on the curvature value. Figure 5 shows the extracted face profile and the absolute values of curvature. Figure 6 gives an example of dynamic time warping of two face profiles from the same person.

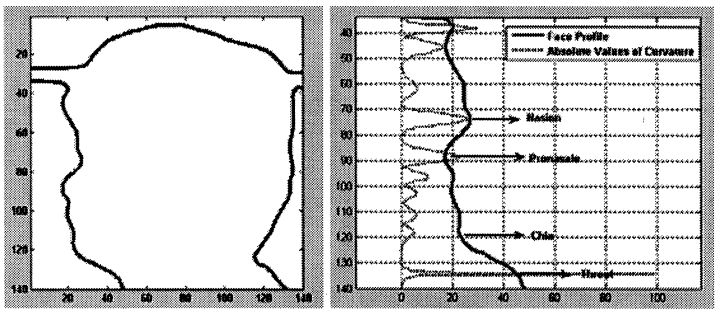

Fig. 5. The extracted face profile and the absolute values of curvature

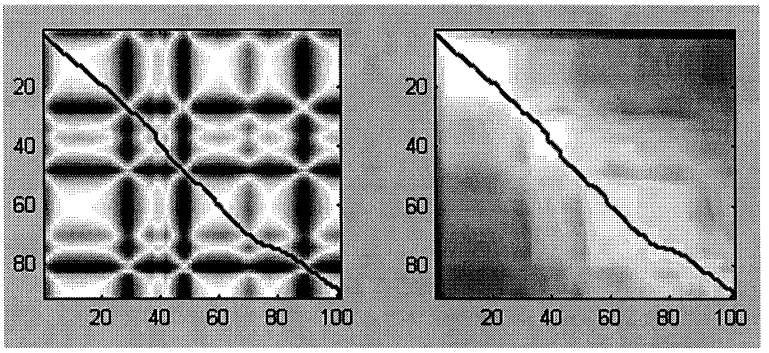

Fig. 6. The similarity matrix (left) and the dynamic programming matrix (right)

From the similarity matrix in Fig. 6, we can see a light stripe (high similarity values) approximately down the leading diagonal. From the dynamic programming matrix in Fig. 6, 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.

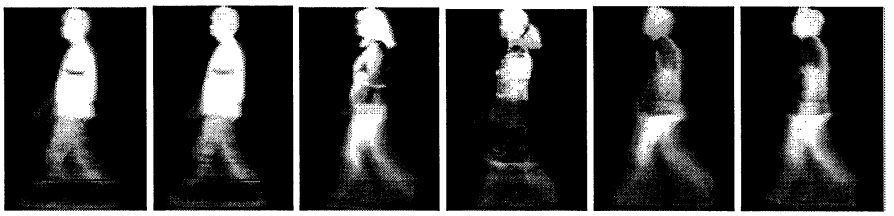

Fig. 7. The Gait Energy Images

2.3 Gait Recognition

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 extraction 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. The obtained time series signal consists of few cycles and lots of noise, which lead to sidelobe effect in the Fourier spectrum. To avoid this problem, we estimate the gait frequency and phase by maximum entropy spectrum estimation.

Gait Representation. Given a preprocessed binary gait silhouette sequence $B(x, y, t)$, the grey-level gait energy image (GEI) is defined as follows [5]:

$$G(x, y) = \frac{1}{N} \sum_{t=1}^{N} B(x, y, t) \tag{5}$$

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), x and y are values in the 2D image coordinate. Figure 7 is some examples of the Gait Energy Images pairs. As expected, it reflects major shapes of silhouettes and their changes over the gait cycle. We refer to it as gait energy image because: (a) each silhouette image is the normalized gait (human walking) area; (b) a pixel within the silhouette in a image means that human walking occurs at this position and this moment; (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. The robustness could be further improved if we discard those pixels with the energy values lower than a threshold. Moreover, with such a 2D template, we do not need to consider the normalized time moment of each frame, and the incurred errors can be therefore avoided.

Direct GEI Matching. One possible approach is recognizing individuals by measuring the similarity between the gallery (training) and probe (testing) templates. Given GEIs of two gait sequences, $G_g(x, y)$ and $G_p(x, y)$, their distance can be measured by calculating their normalized matching error:

$$D(G_g, G_p) = \frac{\sum_{x,y} |G_g(x, y) - G_p(x, y)|}{\sqrt{\sum_{x,y} G_g(x, y) \sum_{x,y} G_p(x, y)}}, \tag{6}$$

where $\sum_{x,y} |G_g(x, y) - G_p(x, y)|$ is the matching error between two GEIs, $\sum_{x,y} G_g(x, y)$ and $\sum_{x,y} G_p(x, y)$ are total energy in two GEIs, respectively.

2.4 Integrating Face Profile and Gait for Recognition at a Distance

Face profile cues and gait cues are considered being integrated by several schemes. Commonly used classifier combination schemes are obtained based on Bayesian Theory, where the representations are assumed conditionally statistically independent. Under different assumptions, there are PRODUCT rule, SUM rule, MAX rule, MIN rule, MEDIAN rule and MAJORITY VOTE rule [6]. We employ SUM rule and PRODUCT rule in our fusion system, with which the similarity scores obtained individually from face profile classifier and gait classifier are combined. Before the similarity scores are combined, it is necessary to map the scores obtained from the different classifiers to the same range of values. Some of the commonly used transformations include linear, logarithmic, exponential and logistic. We use exponential transformation here. The combined similarity score is ranked, which is the result of the fusion system.

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 passes on a smaller set of candidates to the gait classifier. Then the result of the gait classifier is the result of the fusion system.

3 Experimental Results

The data is obtained by Sony DCR-VX1000 digital video camera recorder. We collect 28 video sequences of 14 people walking outside and exposing a side view to the camera, at about 30 frames per second. The shutter speed is 1/60 and the resolution of each frame is 720x480. The distance between people and the instrument is about 7 feet. Each of the persons has two sequences. For 4 of the subjects, the data was collected on two separate days and about 1 months apart. Figure 8 shows the six adjacent video frames of one person.

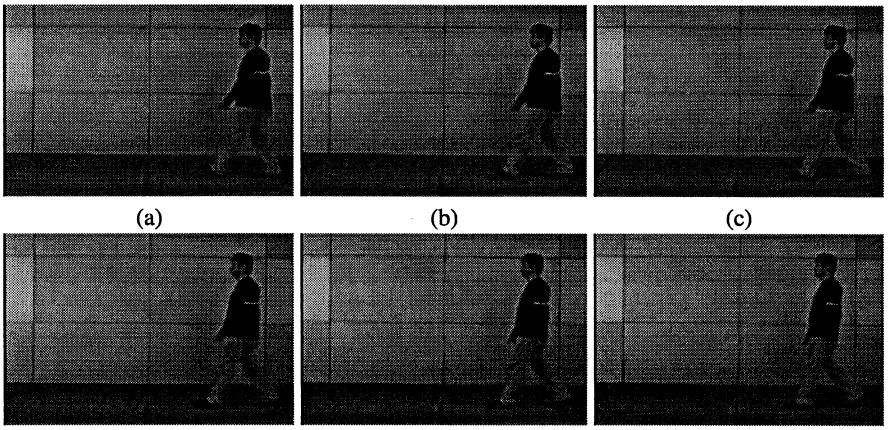

Fig. 8. Six adjacent video frames (a-f)

Table 1. Experimental Results

Combination scheme	Recognition rate		
	Gait	Face profile	Integration
No combination	85.7%	64.3%	
SUM rule			100%
PRODUCT rule			92.9%
Indexing-verification			92.9%

From each sequence, we construct one high-resolution face profile image and one GEI. Since there are two sequences per person, we totally obtain 28 high-resolution face profile images and 28 GEIs for 14 people. Recognition performance is used to evaluate the significance of our method, the quality of extracted features and their impact on identification. The results for our database are shown in Table 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.

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 the clothes of these two people are very different in the training and the testing video sequence, the GEI method can not recognize them correctly. However, the face profiles of these two people don't change so much in the training and the testing sequences. It shows that face profile is a useful cue for the fusion system. On the other hand, since the face profile classifier is comparatively sensitive to the variation of facial expression and

noise, the face profile classifier can not get a good recognition rate by itself. When the gait classifier is combined, the better performance is achieved.

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

4 Conclusions

This paper introduces a practical system combining face profile and gait for human recognition from video. For optimal face profile recognition, we first reconstruct a high-resolution face profile images, using both the spatial and temporal information present in a video sequence. For gait recognition, we use Gait Energy Image (GEI) to characterize human walking properties. Serval schemes are considered for fusion. The experiments show that our system is very promising. Moreover, it is very natural to integrate information of the side face view and the side gait view. However, several important issues that will concern some real-world applications are not addressed in this paper. For example, one problem is how to extract face profile images from video camera automatically and precisely in crowded surveillance applications. Another problem is how to pick up the different frames for the super-resolution algorithm so that the optimal face profile can be reconstructed. These topics will be considered in the future.

References

1. Kale, A., Roychowdhury, A.K., Chellappa, R.: Fusion of gait and face for human identification. Acoustics, Speech, and Signal Processing, 2004. Proceedings. **5** (2004) 901-904
2. Shakhnarovich, G., Lee, L., Darrell, T.: Integrated face and gait recognition from multiple views. Computer Vision and Pattern Recognition, 2001. Proceedings. **1** (2001) 439–446
3. Shakhnarovich, G., Darrell, T.: On probabilistic combination of face and gait cues for identification. Automatic Face and Gesture Recognition, 2002. Proceedings. **5** (2002) 169–174
4. Bhanu, B., Zhou, X.L.: Face recognition from face profile using dynamic time warping. 17th International Conference on Pattern Recogntion. **4** (2004) 499–502
5. Han, Ju, Bhanu, B.:Statistical feature fusion for gait-based human recognition. Computer Vision and Pattern Recognition, 2004. Proceedings. **2** (2004) 842–847
6. Kittler, J., Hatef, M., Duin, R., Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence. **20** (1998) 226–239
7. Zuev, Y., Ivanon, S.: The voting as a way to increase the decision reliability. Foundations of Information/Decision Fusion with Applications to Engineering Problems. **20** (1996) 206–210
8. Tsai, R.Y., Huang, T.S.: Multiframe image resoration and registration. Advances in Computer Vision and Image Processing(T.S. Huang, ed.), JAI Press Inc.. (1984)
9. Borman, S., Stevenson, R.: Spatial resolution enhancement of low-resolution image sequences - a comprehensive review with directions for future research. University of Notre Dame, Tech. Rep.. (1998)

10. Park, S. C., Park, M. K., Kang, M. G.: Super-resolution image reconstruction: A technical overview. IEEE Signal Processing Magazine. **20** (2003) 21–36
11. Irani, M., Peleg, S.: Motion analysis for image enhancement: Resolution, occlusion and transparency. Journal of Visual Communication and Image Representation. **4** (1993) 324–335
12. Irani, M., Peleg, S.: Improving resolution by image registration. CVGIP: Graphical Models and Image Processing. **53** (1991) 231–239