Human Recognition at a Distance

Bir Bhanu
Center for Research in Intelligent Systems
University of California, Riverside, CA 92521, USA

ABSTRACT

Recognizing people at a distance is challenging from various considerations, including sensing, robust processing algorithms, changing environmental conditions and fusing multiple modalities. This paper considers face, side face, gait and ear and their possible fusion for human recognition. It presents an overview of some of the techniques that we have developed for (a) super-resolution-based face recognition in video, (b) gait-based recognition in video, (c) fusion of super-resolved side face and gait in video, (d) ear recognition in color/range images, and (e) fusion performance prediction and validation. It presents various real-world examples to illustrate the ideas and points out the relative merits of the approaches that are discussed

Keywords: Ear Recognition, Face Recognition, Gait Recognition, Multi-modal Fusion, Performance Prediction, Super-Resolution

1. INTRODUCTION

It has been found to be difficult to recognize a person from arbitrary views in changing environmental conditions when a non-cooperative subject is walking at a distance. Some of the challenges include low resolution of the video from single/multiple cameras, changing pose of the subject and uncontrolled illumination conditions. In this paper we address the following problems associated with recognizing people at a distance.

1. Super-resolution and recognition of facial images in video (section 2).
2. Gait-based recognition of humans in video (section 3)
3. Integrated side face and gait recognition in video (section 4)
4. Ear recognition in registered range and color images of the side face (section 5)
5. Performance prediction for sensor fusion (section 6)

2. FACE RECOGNITION IN VIDEO ACQUIRED AT A DISTANCE

There is a growing interest in face recognition and identification for surveillance systems, information security, and access control applications. In many of the above scenarios, the distance between the objects and the cameras is quite large, which makes the quality of video usually low and face images quite small. Low resolution is one of the challenges in video-based face recognition. Enhancing low resolution (LR) images from the video sequence has been studied by many researchers in the past. Traditional approaches in this area first perform tracking in each frame and then use a super-resolution (SR) method for obtaining increased resolution of the imagery. This process does not pass on the benefits of the SR result to the tracking module and inhibits the entire system from reaching its maximum performance potential. However, in real applications, small size images not only make the recognition task more difficult, but also affect the accuracy of face tracking.

We have developed an incremental super-resolution (ISR) technique where SR and tracking are linked together in a closed-loop system. We assume that a 3D generic model is available. The super-resolved texture that is fed back improves the accuracy of pose and illumination estimation, which, in turn, improves the SR result in subsequent frames. We show the comparison between the traditional open loop and our closed-loop framework in Figure 1. Unlike a traditional approach in (a) which treats registration and SR steps separately, our approach in (b) feeds the super-resolved 3D facial texture back to the tracking algorithm, thus increasing the overall quality of tracking and super-resolving the texture over time. In order to compare the difference between the open-loop and our closed-loop approaches, we normalize the illumination to be the same and the pose of SR texture to the frontal pose. The marked points (Figure 1) in

MIPPR 2007: Automatic Target Recognition and Image Analysis; and Multispectral Image Acquisition, edited by Tianxu Zhang, Carl Nardell, Duane Smith, Hangqing Lu, Proc. of SPIE Vol. 6786, 678602, (2007) · 0277-786X/07/$18 · doi: 10.1117/12.774682

the tracking results show the back projection of some mesh vertices using the estimated 3D motion. Since only the first frame is mapped for the 3D texture in the open-loop approach, tracking is slightly off, which in turn gives rise to a poor SR result compared to the closed-loop approach. In our closed-loop super-resolution (SR) approach the fed-back super-resolved texture improves the accuracy of tracking for incoming LR frames. The more accurate tracking, in turn, improves the output of the SR algorithm to acquire better SR texture. Our approach generates SR video by updating the super-resolved texture with the incoming frames.

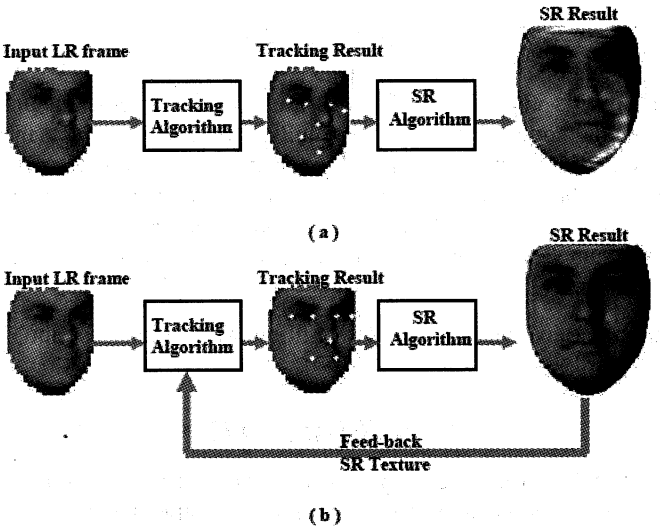

Fig. 1. Open loop vs. closed loop systems for video-based super-resolution from video.

Figure 2 presents the experimental results of comparison between open loop and closed loop systems. Our experimental results demonstrate that our closed-loop approach can significantly improve the accuracy of motion estimation and the quality of SR results compared with traditional open-loop approaches. In various experiments [1] we find that in spite of large changes of pose and lighting, the final super-resolved texture can reach a PSNR in the range of 26-29 dB, the tracking can achieve sub-pixel accuracy with a mean of 0.5 pixel and face recognition can improve over 10-20%. We can treat the entire face as a single unit or treat it in terms of its parts (eyes, lips, eyebrows, and rest of the face) [2]. Since we use 3D face model, our approach can integrate the information over multiple frames from a video sequence as parts of the face become visible from being invisible at the beginning.

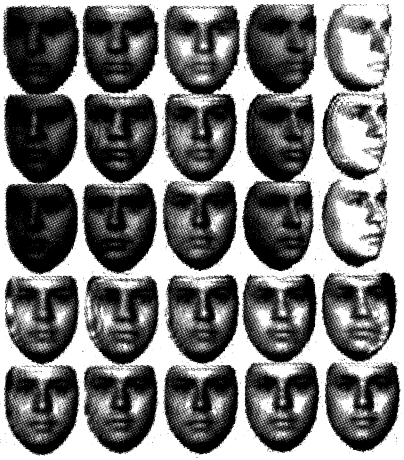

Fig. 2. A comparison of open loop and closed systems for performing super-resolution. First row shows the low resolution input images. Notice the change in pose and illumination. The second row shows the results of open loop super-resolution. The third row shows the results of closed-loop super-resolution. Fourth and fifth row show the illumination and pose normalized super-resolution images obtained by open and closed loop approaches, respectively.

3. GAIT-BASED HUMAN RECOGNITION AT A DISTANCE IN VIDEO

We have developed a representation, called Gait Energy Images (GEI) [3-5] to recognize individuals by their gait as observed in video. GEI is a spatio-temporal compact representation of gait in video. In this representation the entire gait sequence is divided into cycles according to gait frequency and phase information. GEI captures the major shapes of silhouettes and their changes over the gait cycle. Silhouettes in each frame can be obtained using a physically based approach for moving object detection [6]. GEI accounts for human walking at different speeds. It has several advantages over the gait representation of binary silhouette sequence. It is not sensitive to incidental silhouette errors in the individual frames. Moreover, with such a 2D template, we do not need to consider the time moment of each frame, and , therefore, the incurred errors can be avoided.

Given the preprocessed binary gait silhouette sequence in the complete cycle(s), the grey-level gait energy image (GEI) is obtained by averaging the normalized and aligned silhouette images in the gait cycle(s). Figure 3 shows the sample silhouette images in a gait cycle from a person and the right most image is the corresponding GEI. The resolution of each GEI is 300x200.

Various dimensionality reduction techniques such as the Principal Component Analysis (PCA) and subspace methods can be used to develop a compact set of features from GEI for gait-based individual human recognition. For example, we have used PCA and multi-discriminant analysis (MDA) and their various combinations for feature level fusion. A complete set of results on HumanID database, together with their comparison with the state-of-the art algorithms, is given in [3]. In the next section we combine gait with the side face for human recognition.

Fig. 3. Examples of normalized and aligned silhouette images in a gait cycle. The right most image is the gait energy image (GEI).

4. SIDE FACE AND GAIT RECOGNITION IN VIDEO

A fusion system, which combines face and gait cues from a video sequence, is a promising approach to accomplish the task of human recognition at a distance. 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.

The fusion of face and gait is promising in real world applications because of their individual characteristics. Compared with gait, face images are readily interpretable by humans, which allows people to confirm whether a biometrics system is functioning correctly, but the appearance of a face depends on many factors: incident illumination, head pose, facial expressions, moustache/beard, eyeglasses, cosmetics, hair style, weight gain/loss, aging, and so forth. Although gait images can be easily acquired from a distance, the gait recognition is affected by clothes, shoes, carrying status and specific physical condition of an individual. The fusion system is relatively more robust compared with the system that uses only one biometrics. For example, face recognition is more sensitive to low lighting conditions, whereas gait is more reliable under these conditions. Similarly, when the walker is carrying a heavy baggage or he/she is injured, the captured face information may contribute more than gait.

We distinguish a side face from a face problem. A face refers to the outline of the shape of a face as seen from the side. A side face includes not only the outline of the side view of a face, but also the entire side view of eye, nose and mouth, possessing both shape and intensity information. Therefore, a side face has more discriminating power for recognition than a face profile [7]. 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 as the face template [8]. For gait, the Gait Energy Image (GEI), which is used to characterize human walking properties, is generated as the gait template.

We have developed several approaches that integrate information from side face and gait at the feature level and match score level to recognize non-cooperating individuals at a distance. Compared with the abundance of research work

related to fusion at the match score level, fusion at the feature level is a relatively understudied problem because of the difficulties in practice. Multiple modalities may have incompatible feature sets and the relationship between different feature spaces may not be known. Moreover, the concatenated feature vector may lead to the problem of curse of dimensionality and it may contain noisy or redundant data, thus leading to a decrease in the performance of the classifier. However, pattern recognition and computer vision systems that integrate information at an early stage of processing are believed to be more effective than those systems that perform integration at a later stage. Therefore, while it is relatively difficult to achieve in practice, fusion at the feature level has drawn more attention in recent years. Among the existing research work, feature concatenation is the most popular feature level fusion methodology. Some of the schemes perform feature concatenation after dimensionality reduction while others perform feature concatenation before feature selection.

In our new feature level fusion scheme [7-10] we propose to fuse information from side face and gait for human recognition at a distance in a single camera scenario. Multiple Discriminant Analysis (MDA) is carried out after the concatenation of face and gait features using PCA based analysis. This allows the generation of better discriminating features and leads to the improved performance. Face features are extracted from Enhanced Side Face Image (ESFI), which integrates face information over multiple frames in video. Similarly, gait features are extracted from Gait Energy Images (GEI). The concatenation of face and gait features generates better discriminating features for improved recognition performance.

The problem of curse of dimensionality is reduced since the feature vectors are of lower dimension than those in [8]. The problem of the curse of dimensionality is reduced in two ways: (a) PCA is used to transform high dimensional face and gait templates to low dimensional feature space; (b) synthetic features are generated based on all possible combinations of face and gait features from the same video sequence.

The proposed scheme [10] is tested using two comparative data sets to show the effect of changing clothes and face changing over time. Moreover, the proposed feature level fusion (PCA followed by concatenation of features and then MDA) is compared with the match score level fusion and another feature level fusion [9] scheme (PCA, MDA and then the concatenation of features). The experimental results demonstrate that the synthetic features, encoding both side face and gait information, carry more discriminating power than the individual biometrics features. The experimental results show that the proposed feature level fusion scheme [10] is effective for individual recognition in video. It outperforms the previously published fusion schemes at the match score level (Sum and Max rules) and the feature level [8, 9] for face- and gait-based human recognition at a distance in video.

5. EAR RECOGNITION IN REGISTERED COLOR/RANGE IMAGES

Ear, a new class of biometrics, has certain advantages over face recognition at a distance. For example, the ear is rich in features; it is a stable structure that does not change much with age and it does not change its shape with facial expressions. Furthermore, ear is smaller as compared to face and it can be easily captured from a distance without a fully cooperative subject although it can sometimes be hidden with hair, muffler, scarf, and earrings. There are several biometrics techniques using the 2D intensity images. The performance of these techniques is greatly affected by the pose variation and imaging conditions. However, an ear can be imaged in 3D using a range sensor which provides a registered color and range image pair. A range image is relatively insensitive to illuminations and it contains surface shape information related to the anatomical structure, which makes it possible to develop a robust 3D ear biometrics.

We have developed a complete human recognition system using 3D ear biometrics [12]. The system has two key components: 3D ear detection and 3D ear recognition. For ear detection, we propose a two-step approach using the registered 2D color and range images by locating the ear helix and the antihelix parts. In the first step, a skin color classifier is used to isolate the side face in an image by modeling the skin color and nonskin color distributions as a mixture of Gaussians. The edges from the 2D color image are combined with the step edges from the range image to locate regions-of-interest (ROIs) which may contain an ear. In the second step, to locate the ear accurately, the reference 3D ear shape model, which is represented by a set of discrete 3D vertices on the ear helix and the antihelix parts, is adapted to individual ear images by following a new global-to-local registration procedure [13] instead of training an active shape model built from a large set of ears to learn the shape variation. The optimization procedure drives the initial global registration toward the ear helix and the antihelix parts, which results in the one-to-one correspondence of the ear helix and the antihelix between the reference ear shape model and the input image.

The approach for ear detection is followed to build a database of ears that belong to different people. For ear recognition, we have developed two representations: the ear helix/ antihelix representation obtained from the detection algorithm and a new local surface patch representation computed at feature points to estimate the initial rigid transformation between a gallery-probe pair [12]. For the ear helix/antihelix representation, the correspondence of ear

helix and antihelix parts (available from the ear detection algorithm) between a gallery-probe ear pair is established and it is used to compute the initial rigid transformation. For the local surface patch (LSP) representation, a local surface descriptor is characterized by a centroid, a local surface type, and a 2D histogram. The local surface descriptors are computed for the feature points, which are defined as either the local minimum or the local maximum of shape indexes. By comparing the local surface patches for a gallery and a probe image, the potential corresponding local surface patches are established and then filtered by geometric constraints. Based on the filtered correspondences, the initial rigid transformation is estimated. Once this transformation is obtained using either of the two representations, it is then applied to randomly selected control points of the hypothesized gallery ear in the database. A modified Iterative Closest Point (ICP) algorithm is run to improve the transformation, which brings a gallery ear and a probe ear into the best alignment, for every gallery probe pair. The root mean square (RMS) registration error is used as the matching error criterion. The subject in the gallery with the minimum RMS error is declared as the recognized person in the probe.

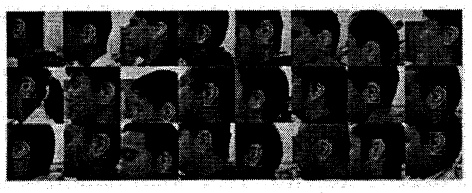

Fig. 4. Ear detection results.

Table 1. Recognition results on UCR and UND datasets using helix/anti-helix and LSP representation..

Dataset	Helix/anti-helix representation					LSP representation				
	Rank-1	Rank-2	Rank-3	Rank-4	Rank-5	Rank-1	Rank-2	Rank-3	Rank-4	Rank-5
$UCR\ \ ES_1(155,155)$	96.77%	98.06%	98.71%	98.71%	98.71%	94.84%	96.77%	96.77%	96.77%	96.77%
$UCR\ \ ES_2(310,592)$	94.43%	96.96%	97.80%	98.31%	98.31%	94.43%	96.96%	97.30%	97.64%	97.80%
$UND(302,302)$	96.03%	96.69%	97.35%	97.68%	98.01%	96.36%	98.01%	98.34%	98.34%	98.34%

The experiments are performed on the data set collected by us (UCR data set) and the University of Notre Dame public data set (UND data set). In the UCR data set, there is no time lapse between the gallery and probe for the same subject, while there is a time lapse of a few weeks (on the average) in the UND data set. UCR data set are captured by Minolta Vivid 300 camera. The camera outputs a range image and its registered color image in less than one second. The range image contains 200x 200 grid points and each grid point has a 3D coordinate (x; y; z) and a set of color (r; g; b) values. During the acquisition, 155 subjects sit on a chair about 0.55-0.75m from the camera in an indoor office environment. The first shot is taken when a subject's left side face is approximately parallel to the image plane; two shots are taken when the subject is asked to rotate his/her head to the left and to the right side within 35 degrees with respect to his/her torso. During this process, there could be some face tilt as well, which is not measured. A total of six images per subject are recorded. In total, 902 shots are used for the experiments since some shots are not properly recorded. Every person has at least four shots. There are three different poses in the collected data: frontal, left, and right. The UND data set are acquired with a Minolta Vivid 910 camera. The camera outputs a 480 x640 range image and its registered color image of the same size. In Collection F, there are 302 subjects with 302 time-lapse gallery-probe pairs. The detection results and shown in Fig. 4 and the recognition results are shown in Table 1. In order to evaluate the proposed surface matching schemes, we perform experiments under two scenarios: 1) One frontal ear of a subject is in the gallery set and another frontal ear of the same subject is in the probe set and 2) two frontal ears of a subject are in the gallery set and the rest of the ear images of the same subject are in the probe set. These two scenarios are denoted as ES1 and ES2, respectively. ES1 is used for testing the performance of the system to recognize ears with the same pose; ES2 is used for testing the performance of the system to recognize ears with pose variations.

6. FUSION PERFORMANCE PREDICTION/VALIDATION

In order to increase the recognition system performance, sensor fusion techniques are widely used today. By fusing different sensors, we may achieve increased accuracy, reduced false alarms and increased range of scenarios for which a system will function correctly. Given the characteristics of the single sensors, how can we find the optimal sensor combination which gives the best recognition performance? The traditional approach is to try all possible combinations of sensors by performing exhaustive experiments to determine the optimal combinations. We have developed a

theoretical approach [14] to predict the sensor fusion performance that allows us to select the optimal sensor combination. In this approach, first, we use the characteristics of each sensor to compute the match score and non-match score distributions which are modeled as a mixture of Gaussians. Second, we decompose the *area under the ROC curve* (*AUROC*) of the fusion system to a set of *AUROC*s which are obtained from the combination of the components from the match score and non-match score distributions. Third, we use an explicit transformation that maps a *ROC* curve to a straight line in 2-D space whose axes are related to the *FAR* and the *Hit* rate. Finally, using this representation, we derive a set of metrics to evaluate the sensor fusion performance and find the optimal sensor combination. By using this approach, we can determine the optimal sensor combination by computing the metrics instead of performing the exhaustive experiments. We have also developed another prediction model which is based on the likelihood ratio to predict the sensor fusion performance. We derive the Fisher measurement for the sensor fusion system to predict the system performance. In this approach, we model the match score and non-match score as the single Gaussian distributions. We verify our prediction approach on the multi-modal XM2VTS database, NIST-4 fingerprint database, ear database and video databases. We use Fowlkes and Mallows index to evaluate the degree of the agreement between the fusion performance evaluation and prediction. The experimental results show that our prediction approach can predict the sensor fusion performance effectively.

7. CONCLUSIONS

Video-based human recognition at a distance remains a challenging problem for individual and multi-modal biometrics systems based on face, gait, side face, and ear. In this paper we have briefly described our work in this area. These biometrics can be used in both the low and high security scenarios and are of interest in networked applications.

REFERENCES

1. J. Yu, B. Bhanu, Y. Xu and A.K. Roy-Chowdhury, "Super-resolved facial texture under changing pose and illumination," *International Conference on Image Processing*, (Sept. 2007).
2. J. Yu and B. Bhanu, "Super-resolution restoration of facial images in video," *International Conference on Pattern Recognition*, (Aug. 2006).
3. J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2), 316-322, (February 2006).
4. J. Han and B. Bhanu, "Moving human detection by EO and IR sensor fusion," *Pattern Recognition*, 40(6): 1771-1784 (2007)
5. J. Han and B. Bhanu, "Performance prediction for individual recognition by gait," *Pattern Recognition Letters*, 26(5), 615-624, (April 2005).
6. S. Nadimi and B. Bhanu, "Physical models for moving shadow and object detection in video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8), 1079-1087, (August 2004).
7. X. Zhou, B. Bhanu and J. Han, "Human recognition at a distance in video by integrating face profile and gait," *Proceedings International Conference on Audio- and Video-based Biometric Person Authentication*, pp. 533-543, 2005.
8. X. Zhou and B. Bhanu, "Integrating face and gait for human recognition at a distance in Video," *IEEE Transactions on Systems, Man and Cybernetics* Part B, Special Issue on Biometrics, (Oct. 2007).
9. X. Zhou and B. Bhanu, "Feature fusion of face and gait for human recognition at a distance in video," *International Conference on Pattern Recognition*, Hong Kong, China, Aug. 21-24, 2006.
10. X. Zhou and B. Bhanu, "Feature fusion for video-based human identification," *Pattern Recognition*, Special issue on Biometrics, in Press, (2007).
11. H. Chen and B. Bhanu," Human ear recognition in 3D," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Special Issue on Biometrics, 29(4), (April 2007).
12. H. Chen and B. Bhanu, "3D free-form object recognition in range images using local surface patches," *Pattern Recognition Letters,* in Press, (2007).
13. H. Chen and B. Bhanu, "Global to local non-rigid shape registration," *International Conference on Pattern Recognition*, Hong Kong, China, Aug. 21-24, 2006.
14. R. Wang and B. Bhanu, "On the performance prediction for sensor fusion," *IEEE Conference on Computer Vision and Pattern Recognition.*, (June 2007).
15. R. Wang and B. Bhanu, "Performance prediction for multimodal biometrics," *International Conference on Pattern Recognition*, (Aug. 2006).

PROCEEDINGS OF SPIE

MIPPR 2007

Automatic Target Recognition and Image Analysis; and Multispectral Image Acquisition

Tianxu Zhang
Carl Nardell
Duane Smith
Hangqing Lu
Editors

15–17 November 2007
Wuhan, China

Sponsored by
State Key Laboratory for Multi-spectral Information Processing Technologies (China)
Chinese Education Ministry Key Laboratory for Image Processing and Intelligence Control (China)
Huazhong University of Science and Technology (China)

Technical Sponsor
SPIE

Volume 6786
Part One of Two Parts

SPIE