An Iterative Method for Multimodal Biometric Face Recognition Using Speech Signal

M. Nageshkumar* and M.N. ShanmukhaSwamy

Department of Electronics and Communication., J.S.S. Research Foundation,
University of Mysore., Mysore-06
nageshkumar79m@gmail.com, mnsjce@gmail.com

Abstract. In recent years much advancement have been made in face recognition techniques to cater to the challenges such as pose, expression, illumination, aging and disguise. However, due to advances in technology, there are new emerging challenges for which the performance of face recognition systems degrades and plastic/cosmetic surgery is one of them. In this paper we comment on the effect of plastic surgery face image in multimodal biometric face recognition using speech signal. Speaker identity is correlated with the physiological and behavioral characteristics of the speaker. Selecting the most effective fusion techniques depends on operational issues such as accuracy requirements, availability of training data, and the validity of simplifying assumptions.

Keywords: Multimodal biometric system, plastic surgery face image, speech signal and matching level fusion.

1 Introduction

Human biometric characteristics are unique, so it can hardly be duplicated [1]. Such information includes: facial, speech, hands, body, fingerprints, and gesture to name a few. Face detection and recognition techniques are proven to be more popular than other biometric features based on efficiency and convenience [2, 3]. Face authentication has become a potential a research field related to face recognition. Face recognition differs from face authentication because the former has to determine the identity of an object, while the latter needs to verify the claimed identity of a user. Speech [4] is one of the basic communications, which is better than other methods in the sense of efficiency and convenience. Each a single biometric information, however, has its own limitation. For this reason, we present a multimodal biometric verification/identification method to reduce false acceptance rate (FAR) and false rejection rate (FRR) in real-time.

A multimodal biometric face recognition is a well studied problem in which several approaches have been proposed to address the challenges of illumination [9,10], pose [11,12,13], expression [10], aging [14,15] and disguise [16,17], the growing popularity of plastic surgery introduces new challenges in designing future face

* Corresponding author.

N. Meghanathan et al. (Eds.): CCSIT 2011, Part I, CCIS 131, pp. 298–306, 2011.

recognition systems. Since these procedures modify both the shape and texture of facial features to varying degrees, it is difficult to find the correlation between pre and post surgery facial geometry. Due to the sensitive nature of the process and the privacy issues involved, it is extremely difficult to prepare a face database that contains images before and after surgery. After surgery, the geometric relationship between facial features changes and there is no technique to detect and measure such type of alterations.

The main aim of this paper is to add a new dimension to face recognition by using speech signal and discussing this challenge and systematically evaluating the performance of existing faces recognition algorithms on a database that contains face images before and after surgery.

Related Work

Brunelli and Falavigna [18] used hyperbolic tangent (tanh) for normalization and weighted geometric average for fusion of voice and face biometrics. Kittler [19,23] have experimented with several fusion techniques for face and voice biometrics, including sum, product, minimum, median, and maximum rules and they have found that the sum rule outperformed others.

Hong and Jain [20] proposed an identification system based on face and fingerprint, where fingerprint matching is applied after pruning the database via face matching. Ben-Yacoub [21, 24, 25] considered several fusion strategies, such as support vector machines, tree classifiers and multi-layer perceptions, for face and voice biometrics. Ross and Jain [22] combined face, fingerprint and hand geometry biometrics with sum, decision tree and linear discriminant-based methods. The authors report that sum rule outperforms others.

The rest of this paper is organized as follows. Section 2 presents the proposed Diagonal PCA method for face feature extraction. Section 3 presents the speech feature extraction method. Also section 4 presents the fusion at matching score level. Section 5 reports on the experimental results. Finally, Section 6 concludes.

2 Face Feature Extraction

Our motivation for developing the Diagonal PCA method originates from an essential observation on the recently proposed 2DPCA [28]. That is, 2DPCA can be seen as the row-based PCA [26, 27], which has been pointed out in [29]. So 2DPCA only reflects the information between rows, which implies some structure information (e.g. regions of a face like eyes, nose, etc.) cannot be uncovered by it. We attempt to solve that problem by transforming the original face images into corresponding *diagonal face images*. Because the rows (columns) in the transformed diagonal face images simultaneously integrate the information of rows and columns in original images, it can reflect both information between rows and those between columns. Through the entanglement of row and column information, it is expected that Diagonal PCA may find some useful block or structure information for recognition in original images. Suppose that there are M training face images, denoted by m by n matrices $A_k (k = 1, 2, ...M)$.

For each training face image, define the corresponding *diagonal face image* as follows:

1) If the height m is equal to or smaller than the width n, use the method illustrated in Figure.1 to generate the diagonal image B for the original image A.

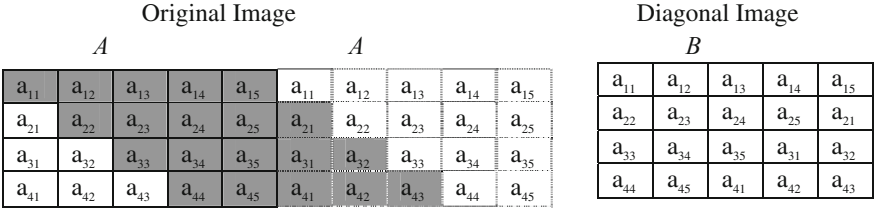

Fig. 1. Illustration for deriving the diagonal face images

Without loss of generalization, assume that the width n is no smaller than the height m. For each training face image A_k, derive the corresponding diagonal face B_k using the method illustrated in Figure.1 Note that B_k *s* is of the same size of A_k *s*.

2) If the height m is bigger than the width n, use the method illustrated in Figure.2 to generate the diagonal image B for the original image A.

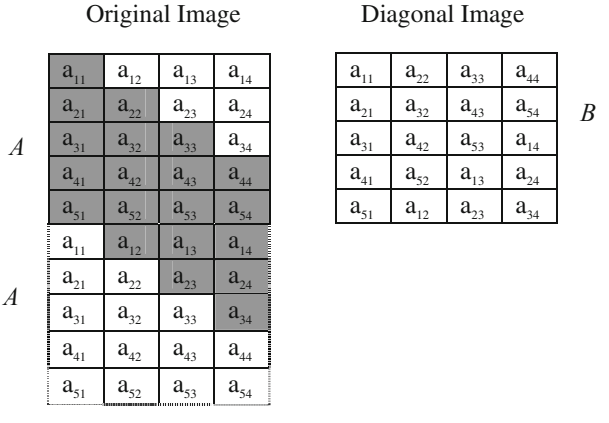

Fig. 2. Illustration for deriving the diagonal face images

Based on the diagonal faces, define the *diagonal covariance matrix* as

$$G = \frac{1}{M}\sum_{k=1}^{M}\left(B_k - \overline{B}\right)^T\left(B_k - \overline{B}\right) \tag{1}$$

where $\overline{B} = \frac{1}{M}\sum_k B_k$ is the mean diagonal face image, according to eq. (1), the projective vectors X_1.....X_d can be obtained by computing the d eigenvectors corresponding

to the *d* biggest eigenvalues of G. Since the size of G is only *n* by *n*, computing its eigenvectors can be efficient. Let $X=[X_1.....X_d]$ denote the projective matrix, projecting training faces Ak s onto X, yielding *m* by *d feature matrices*

$$C_k = A_k X \tag{2}$$

Given a test face image A, first use eq. (2) to get the feature matrix $C=AX$, then a nearest neighbor classifier can be used for classification. Here the distance between C and Ck is defined as:

$$d(C, C_k) = \|C - C_k\| = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{d}\left(C^{(i,j)} - C_k{}^{(i,j)}\right)^2} \tag{3}$$

3 Speech Feature Extraction

The objective of voice recognition is to determine which speaker is present based on the individual's utterance [30].Several techniques have been proposed for reducing the mismatch between the testing and training environments. Many of these methods operate either in spectral [31, 32], or in cepstral domain [33]. First, the speech samples are processed using MFCC to produce voice features. After that, the coefficient of voice features can go trough DTW to select the pattern that matches the database and input frame in order to minimize the resulting error between them.

Text-dependent methods are usually based on template-matching techniques. In this approach, the input utterance is represented by a sequence of feature vectors, generally short-term spectral feature vectors. The time axes of the input utterance and each reference template or reference model of the registered speakers are aligned using a dynamic time warping (DTW) algorithm and the degree of similarity between them, accumulated from the beginning to the end of the utterance, is calculated.

A voice analysis is done after taking an input through microphone from a user. The design of the system involves manipulation of the input audio signal. At different levels, different operations are performed on the input signal such as Pre-emphasis, Framing, Windowing, Mel Cepstrum analysis and Recognition (Matching) of the spoken word.

3.1 Feature Extraction Using MFCC

***Step 1*:** Pre–emphasis
This step processes the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.

$$Y[n] = X[n] - 0.95\,X[n-1] \tag{4}$$

Lets consider a = 0.95, which make 95% of any one sample is presumed to originate from previous sample.

Step 2: Framing
The voice signal is divided into frames of N samples. Adjacent frames are being separated by M (M<N). Typical values used are M = 100 and N= 256.

Step 3: Hamming windowing
Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines. The Hamming window equation is given as:

If the window is defined as W (n), $0 \leq n \leq N\text{-}1$

Where- N = number of samples in each frame
$Y[n]$ = Output signal
$X(n)$ = input signal
$W(n)$ = Hamming window, then the result of windowing signal is shown below:

$$Y(n) = X(n) \times W(n) \tag{5}$$

$$w(n) = 0.54 - 0.46\,COS\left(\frac{2\pi n}{N-1}\right) \qquad 0 \leq n \leq N-1 \tag{6}$$

Step 4: Fast Fourier Transform
To convert each frame of N samples from time domain into frequency domain. The Fourier Transform is to convert the convolution of the glottal pulse $U[n]$ and the vocal tract impulse response $H[n]$ in the time domain. This statement supports the equation below:

$$Y(w) = FFT\,[\,h(t) * X(t)] = H(w) * X(w) \tag{7}$$

If $X(w)$, $H(w)$ and $Y(w)$ are the Fourier Transform of $X(t)$, $H(t)$ and $Y(t)$ respectively.

Step 5: Mel Filter Bank Processing
The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. To compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters [7, 8]. Then, each filter output is the sum of its filtered spectral components. After that the following equation is used to compute the Mel for given frequency f in HZ:

$$F(Mel) = [\,2595 * log\ 10\ [1+f]\ 700\,] \tag{8}$$

Feature Matching Using DTW

The principle of DTW is to compare two dynamic patterns and measure its similarity by calculating a minimum distance between them.

Suppose we have two time series P and Q, of length n and m respectively, where:

$$P = p_1, p_2, \ldots . p_i, \ldots \ldots . p_n \tag{9}$$

$$P\ Q = q_1, q_2, \ldots . q_i, \ldots \ldots . q_m \tag{10}$$

To align two sequences using DTW, an *n*-by-*m* matrix where the (i^{th}, j^{th}) element of the matrix contains the distance $d(p_i, q_i)$ between the two points p_i and q_i is constructed. Then, the absolute distance between the values of two sequences is calculated using the Euclidean distance computation:

$$d(p_i, q_i) = (p_i - q_i)^2 \tag{11}$$

Each matrix element (i, j) corresponds to the alignment between the points p_i and q_i. Then, accumulated distance is measured by:

$$D(i, j) = min\,[\,D(i-1, j-1), D(i-1, j), D(i, j-1)] + d(i,j) \tag{12}$$

4 Fusion at the Matching Score Level

In the context of verification, there are two approaches for consolidating the scores obtained from different matchers. One approach is to formulate it as a classification problem, where a feature vector is constructed using the matching scores output by the individual matchers; this feature vector is then classified into one of the two classes: "Genuine user" or "Impostor". In the combination approach, the individual matching scores are combined to generate a single scalar score, which is then used to make the final decision. Now consider a multimodal biometric verification system that utilizes the combination approach to fusion at the match score level. The theoretical framework developed by Kittler [34] can be applied to this system only if the output of each modality is of the form *P (genuine|Z)* i.e., the posteriori probability of user being "genuine" given the input biometric sample *Z*. In practice, most biometric systems output a matching score *s*, and Verlinde [35] have proposed that the matching score *s* is related to *P (genuine|Z)* as follows:

$$S = f\{P(genuine \mid Z)\} + \eta(Z) \tag{13}$$

where *f* is a monotonic function and η is the error made by the biometric system that depends on the input biometric sample *Z*. This error could be due to the noise introduced by the sensor during the acquisition of the biometric signal and the errors made by the feature extraction and matching processes.If we assume that η is zero, it is reasonable to approximate *P(genuine|Z)* by *P(genuine|s)*. In this case, the problem reduces to computing *P(genuine|s)* and this requires estimating the conditional densities *P(s|genuine)* and *P(s|impostor)*. The probability of the score being that of a genuine user was then computed as,

$$P(genuine \mid s) = \frac{p(s \mid genuine)}{p(s \mid genuine) + p(s \mid imposter)} \tag{14}$$

5 Experimental Results

To evaluate the performance of face recognition algorithms in such an application scenario, the plastic surgery database is partitioned into two groups: training database and testing database. This partition ensures that the verification is performed on

unseen images. The train-test partitioning is repeated again and again by computing the false rejection rates (FRR) over these trials at different false accept rate (FAR). The verification accuracy is computed at 5.26% FAR. The experimental result for the recognition rate using the proposed method is summarized in Table 1. An experimental result of FAR given in Table 1 corresponds to 5.26%. In this case, the FAR can accept a person out of 120. Table 2 shows the result of the recognition rate and FAR for the proposed method.

Table 1. Verification rates of face and speech

Test Database	Verification Rates (%)	FAR (%)
Face	88.52	11.48
Speech	92.37	7.63

Table 2. Verification rate of the proposed method

Test Database	Verification Rates (%)	FAR (%)
Fusion of Face & Speech	94.74	5.26

6 Conclusion

In this paper, we present a multimodal biometric human identification method using combined plastic surgery face image and speech information in order to improve the problem of multimodal biometric face recognition system. Current face recognition algorithms mainly focus on handling pose, expression, illumination, aging and disguise. This paper formally introduces plastic surgery as another major challenge for face recognition algorithms using speech signal. Based on the results, we believe that more research is required to design optimal face recognition algorithms that can account for the challenges due to plastic surgery. The procedures can significantly change the facial regions both locally and globally, altering the appearance, facial features and texture. Existing face recognition algorithms generally rely on this information and any variation can affect the multimodal biometric recognition performance.

References

1. Kong, S., Heo, J., Abidi, B., Paik, J., Abidi, M.: Recent advances in visual and infrared face recognition - A review. Computer Vision and Image Understanding 97(1), 103-135, 1077–3142 (2005)
2. Kriegman, D., Yang, M., Ahuja, N.: Detecting faces in images: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(1), 34-58, 0162–8828 (2002)
3. Liu, X., Chen, T., Kumar, V.: On modeling variations for face authentication. In: Proceedings of International Conference Automatic Face and Gesture Recognition, pp. 369–374 (May 2002)

4. Gu, Y., Thomas, T.: A hybrid score measurement for HMM-based speaker verification. In: Proceedings of IEEE International Conference Acoustics, Speech, and Signal Processing, vol. 1, pp. 317–320 (1999)
5. Rowley, H., Baluja, S., Kanade, T.: Neural network-based face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(1), 203-208, 0162–8828 (1998)
6. Osuna, E., Freund, R., Girosi, F.: Training support vector machines: an application to face detection. In: Proceeding of IEEE Conference Computer Vision and Pattern Recognition, pp. 130–136 (1997)
7. Samaria, F., Young, S.: HMM based architecture for face identification. Image and Vision Computing 12(8), 537-543, 262–8856 (1994)
8. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs fisherfaces: recognition using class specification linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), 711-720, 162–8828 (1997)
9. Li, S., Chu, R., Liao, S., Zhang, L.: Illumination invariant face recognition using near-infrared images. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(4), 627–639 (2007)
10. Singh, R., Vatsa, M., Noore, A.: Improving verification accuracy by synthesis of locally enhanced biometric images and deformable model. Signal Processing 87(11), 2746–2764 (2007)
11. Blanz, V., Romdhami, S., Vetter, T.: Face identification across different poses and illuminations with a 3d morphable model. In: Proceedings of International Conference on Automatic Face and Gesture Recognition, pp. 202–207 (2002)
12. Liu, X., Chen, T.: Pose-robust face recognition using geometry assisted probabilistic modeling. In: Proceedings of International Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 502–509 (2005)
13. Singh, R., Vatsa, M., Ross, A., Noore, A.: A mosaicing scheme for pose-invariant face recognition. IEEE Transactions on Systems, Man and Cybernetics - Part B 37(5), 1212–1225 (2007)
14. Lanitis, A., Taylor, C.J., Cootes, T.F.: Toward automatic simulation of aging effects on face images. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(4), 442–450 (2002)
15. Ramanathan, N., Chellappa, R.: Face verification across age progression. IEEE Transactions on Image Processing 15(11), 3349–3362 (2006)
16. Ramanathan, N., Chowdhury, A.R., Chellappa, R.: Facial similarity across age, disguise, illumination and pose. In: Proceedings of International Conference on Image Processing, pp. 1999–2002 (2004)
17. Singh, R., Vatsa, M., Noore, A.: Face recognition with disguise and single gallery images. Image and Vision Computing 27(3), 245–257 (2009)
18. Brunelli, R., Falavigna, D.: Person Identification Using Multiple Cues. IEEE Trans. PAMI 17(10), 955–966 (1995)
19. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On Combining Classifiers. IEEE Trans. PAMI 20(3), 226–239 (1998)
20. Hong, L., Jain, A.K.: Integrating Faces and Fingerprints for Personal Identificaion. IEEE Trans. PAMI 20(12), 1295–1307 (1998)
21. Ben-Yacoub, S., Abdeljaoued, Y., Mayoraz, E.: Fusion of Face and Speech Data for Person Identity Verification. IEEE Trans. N Networks 10(5) (1999)
22. Ross, A., Jain, A.K.: Information Fusion in Biometrics. Pattern Recognition Letters 24(13), 2115–2125 (2003)

23. Kittler, J., Hatef, M., Duin, R., Matas, J.: On Combining Classifiers. IEEE Trans on Pattern Analysis and Machine Intelligence 20(3) (March 1998)
24. Ben-Yacoub, S., Abdeljaoued, S., Mayoraz, E.: Fusion of Face and Speech Data for Person Identity Verification (1999)
25. Fierrez-Aguilar, J., Ortega-Garcia, J., Garcia-Romero, D., Gonzalez-Rodriguez, J.: A comparative evaluation of fusion strategies for multimodal biometric verification. In: Kittler, J., Nixon, M.S. (eds.) AVBPA 2003. LNCS, vol. 2688, pp. 830–837. Springer, Heidelberg (2003)
26. Chen, S.C., Zhu, Y.L., Zhang, D.Q., Yang, J.Y.: Feature extraction approaches baseed on matrix pattern: MatPCA & MatFLDA. Pattern Recognition Letters 26(8) (2005)
27. Turk, M., Pentland, A.: Eigenfaces for recog. J. Cognitive Neuroscience (1991)
28. Yang, J., Zhang, D., Frangi, A.F., Yang, J.Y.: Two-dimensional PCA: a new approach to appearance-based face representation and recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence 26(1), 131–137 (2004)
29. Zhang, D.Q., Chen, S.C., Liu, J.: Representing image matrices: Eigenimages vs. Eigenvectors. In: Wang, J., Liao, X.-F., Yi, Z. (eds.) ISNN 2005. LNCS, vol. 3497, pp. 659–664. Springer, Heidelberg (2005)
30. Yee, C.S., Ahmad, A.M.: Malay Language Text Independent Speaker Vertification using NN-MLP classsifier with MFCC (2008)
31. Lockwood, P., Boudy, J.: Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the Projection, for Robust Speech Reco (1992)
32. Rosenberg, A., Lee, C.-H., Soong, F.: Cepstral Channel Normalization Techniques for HMM Based Speaker Verification (1994)
33. Jackson, P.: Features extraction 1.ppt, University of Surrey, GU2 & 7XH
34. Kittler, J., Hatef, M., Duin, R.P., Matas, J.G.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. 20(3), 226–239 (1998)
35. Verlinde, P., Druyts, P., Cholet, G., Acheroy, M.: Applying Bayes based classifiers for decision fusion in a multimodal identity verification system. In: Proceedings of International Symposium on Pattern Recognition In Memoriam Pierre Devijver, Brussels, Belgium (1999)