

Multi-Modal Human Verification Using Face and Speech

Changhan Park¹ and Joonki Paik²

¹*Advanced Technology R&D Center, Samsung Thales Co., Ltd.*, ²*Graduate School of
Advanced Imaging Science, Multimedia, and Film Chung-Ang University, Seoul
Korea*

1. Introduction

Human biometric characteristics are unique, so it can hardly be duplicated (Kong et al. 2005). 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 (Kriegman et al. 2002; Liu et al. 2002). It can also use a low-cost personal computer (PC) camera instead of expensive equipments, and require minimal user interface. 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 (Gu and Thomas 1999) 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 method to reduce false acceptance rate (FAR) and false rejection rate (FRR) in real-time.

There have been many approaches for extracting meaningful features. Those include principal component analysis (PCA) (Rowley et al. 1998), neural networks (NN) (Rowley et al. 1998), support vector machines (SVM) (Osuna et al. 1997), hidden markov models (HMM) (Samaria and Young 1994), and linear discriminant analysis (LDA) (Belhumeur et al. 1997). In this chapter, we use the PCA algorithm with unsupervised learning to extract the face feature. We also use the HMM algorithm for extracting speech feature with supervised learning.

This chapter is organized as follows: Section 2 and 3 describe feature extraction of face and speech using the PCA and HMM algorithms, respectively. Section 4 presents the design and structure of the proposed system. Section 5 presents experimental, and Section 6 concludes the paper with future research topics.

2. Face Extraction and Recognition

In this section, the proposed face extraction and recognition method will be presented. The proposed method can deal with both gray and color images. Depending on the type of images, an additional preprocessing step may be included so that facial features can be detected more easily.

2.1 Face feature extraction and recognition

The proposed face feature extraction and recognition method is shown in Figure 1. The proposed method makes a new edge image using a 13×9 template in the face image. It can also estimate the face poses and normalize the size of detected face to 60×60 . The normalized image is stored in multimodal database, and it trains the PCA module. The face recognition module distinguishes an input image from trained images.

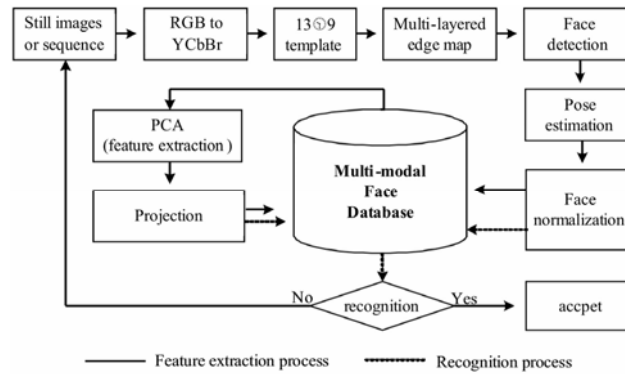


Figure 1. Face feature extraction and recognition process

2.2 Face detection and building database using multi-layered relative edge map

In order to detect a face region and estimate face elements, we use the multi-layered relative edge map which can provide better result than just color-based methods (Kim et al. 2004). Such directional blob template can be determined according to the face size. More specifically, the template is defined so that the horizontal axis is longer than the vertical axis as shown in Figure 2(a). The central pixel of a template in a $W \times H$ image is defined as $P_c = (x_c, y_c)$, which is created by averaging a 3×3 region. By using a $w_{ff} \times h_{ff}$ directional template for face components, the average intensity $\overline{I_{Dir}}$ of 8-neighborhood pixels is calculated on the central pixel, P_c . As a result, $\overline{I_c}$, the brightness value at P_c , and the brightness difference value can be obtained. The principal direction, $\overline{d_{pr}}$, and its magnitude, $|\overline{d_{pr}}|$, are also determined along the direction including the biggest brightness difference as shown in Figure 2(b).

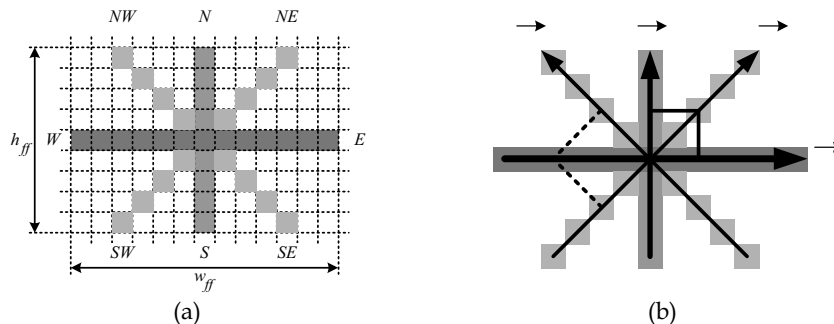


Figure 2. (a) Directional template (b) New direction for edge map

Figure 3 shows the result of face separation by using the multi-layered relative edge map (MLREM) and with this result we make the face database.

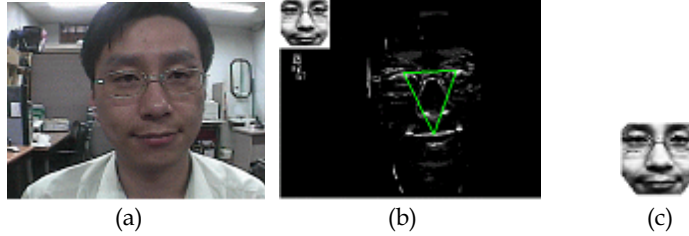


Figure 3. (a) An input image, (b) the correspondingly created MLREM, and (c) the normalized database (60×60)

2.3 Unsupervised PCA and singular value decomposition (SVD)

In the process of PCA for pose estimation we compute covariance matrix C and its eigenvectors from training sets. Let x_1, x_2, \dots, x_N be N training face vectors. By definition, C can then be estimated as (Zhang et al. 1997),

$$C = E[XX^T] = \frac{1}{N} \sum_{k=1}^N X_k X_k^T. \quad (1)$$

The training data set are packed into the following matrix

$$X = [x_1, x_2, \dots, x_N]. \quad (2)$$

The estimate of C can be approximately written as

$$C = \frac{1}{N} XX^T. \quad (3)$$

To estimate the eigenvectors of C , we only need to find the eigenvectors of XX^T . Even for images of moderate size, however, this is computational by complex. From the fundamental linear algebra (Sirivich and Kirby 1987), the eigenvectors of XX^T can be found from eigenvectors of $X^T X$, which are much easier to obtain. Suppose the rank of X is r , $r \leq N$. X has a SVD such as

$$X = \sum_{k=1}^r \sqrt{\lambda_k} u_k v_k^T, \quad (4)$$

where $\sqrt{\lambda_k}$, u_k , and v_k respectively represent, singular values, left, and right singular vectors of X . u_k and v_k have the following relationship.

$$u_k = \frac{1}{\sqrt{\lambda_k}} X v_k. \quad (5)$$

Hence, we can easily find eigenface u_k after finding v_k . Recognized face classified using

$d = \sum_{i=1}^m (r_i - t_i)^2$, where r_i and t_i represent input pattern, pattern of train face, respectively.

3. Speech Analysis and Feature Extraction

Speech recognition is classified into two categories in the sense of feature extraction method. One is to extract a linguistic information in speech signal, and the other is to extract an eigen specific of a speaker from speech signal (Rabiner and Juang 1998). The former performs extraction using the Mel-frequency cepstral coefficient (MFCC) based on the sense of hearing for human, and the latter extracts it with using the linear predictive coefficient (LPC) based on the sense of human speech. We adopt the latter because an individual has its own sense of speech. The LPC processing for speech recognition is shown as Figure 4. A simulation result of LPC in the proposed method is shown as Figure 5.

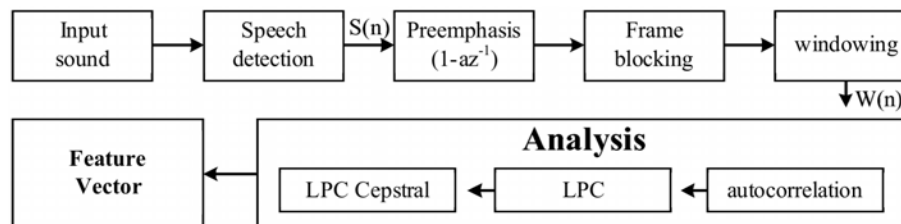


Figure 4. LPC processing for speech recognition

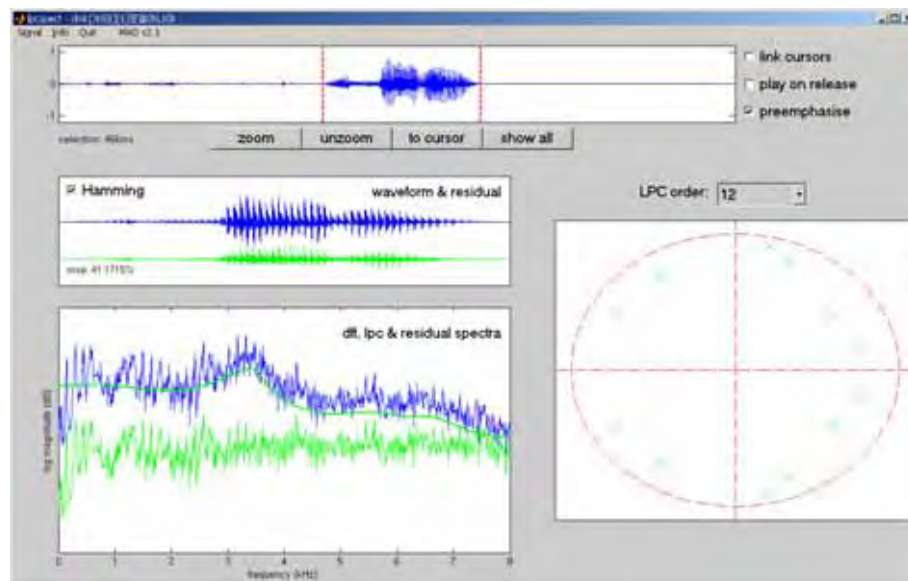


Figure 5. A simulation of LPC coefficient of 12th for Korean (open door)

3.1 HMM for speech recognition and verification

Speech verification calculates the cumulative distances with reference pattern when the test pattern is input. The reference patterns should be made in advance, and it can represent each speaker. This is classified in the pattern matching method that recognizes the pattern with calculated minimal cumulative distances and HMM. The HMM measures similarity

with input pattern after modeling the speech signal statistically by extracting the feature from various speech waveforms. Training and verification for speech are shown in Figure 6. And the proposed method can solve following three problems:

- (i) Evaluation problem: Given an observation sequence $O = \{o_1, o_2, \dots, o_T\}$ and the model $\lambda = (A, B, \pi)$, (where, A represents transition probability, B output probability, and π initial probability), how to calculate $P(O | \lambda)$ - (it can be solved by using forward and backward algorithm.)
- (ii) Learning problem: How to estimate the model parameter given $O = \{o_1, o_2, \dots, o_T\}$ - (It can be solved by using Baum-Welch re-estimation.)
- (iii) Decoding (recognition) problem: Given a model, how to get the best state sequence $q = \{q_1, q_2, \dots, q_i\}$ of $O = \{o_1, o_2, \dots, o_T\}$, where q represents the state sequence of model, t time. - (It can be solved by using the Viterbi algorithm.), where O represents specific vector for each frame.

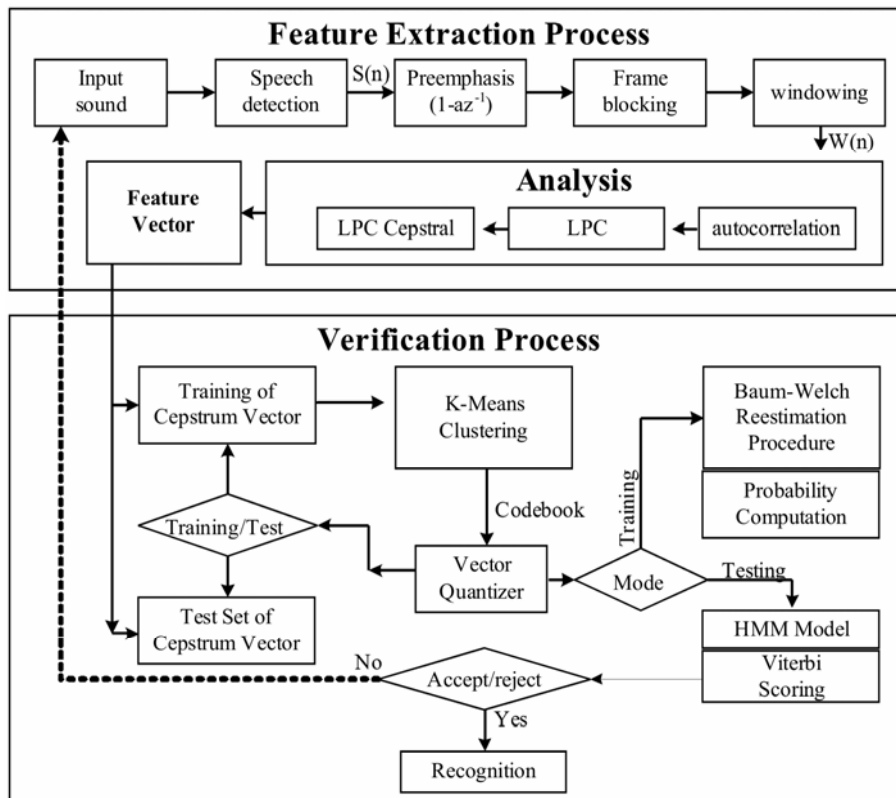


Figure 6. Feature extraction and verification for speech

4. Proposed Multimodal Biometric System

The proposed multimodal biometric recognition technique, can solve the fundamental limitations inherit to single biometric verification. The proposed verification system consists of the input, the learning, and the verification module. The input image of size 320×240 comes into the system in real-time together with the speech. In the learning module, the face image is trained under the PCA framework, and the speech is trained with HMM. Feature extraction is also accomplished in the learning module. The verification module validates the recognized data from the image and speech by using fuzzy logic. Personal information made is saved in the form of a code book, and used for verification and rejection.

4.1 Personal verification using multimodal biometric

In this subsection, we present a personal verification method as shown in Figure 6. The proposed method first detects the face area in an input image. The face verification module compares the detected face with the pre-stored code book of personal information. The speech verification module extracts and recognizes the end-point of speech, and authenticates it after comparing with the code book. Decision processes of face and speech use the proposed fuzzy logic algorithm. If the face and speech verification results coincide, there is no further processing. Otherwise the fuzzy logic is used to solve the mismatch problem. Therefore, if the face and speech is same to the personal information of the code book verification is accepted. Otherwise, it is rejected. The entire verification process is shown in Figure 7.

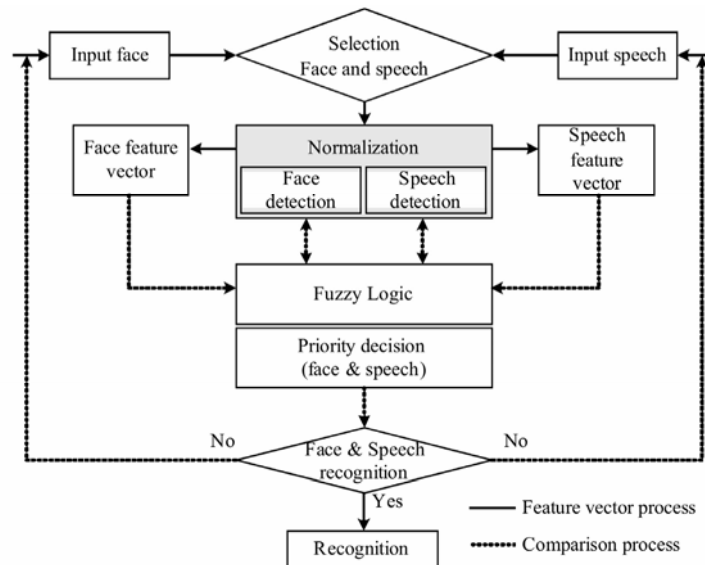


Figure 7. The entire verification process

4.2 Code book of personal face and speech information

In this subsection, the proposed personal information code book is described as shown in Figure 8. The face feature extraction block is trained by using the PCA algorithm with ten

different images per single person. Each an individual probability information projects the data to the original image. Figure 9 shows a set of registered face images. The speech feature extraction block is trained by using the HMM algorithm with ten iterations per single person.

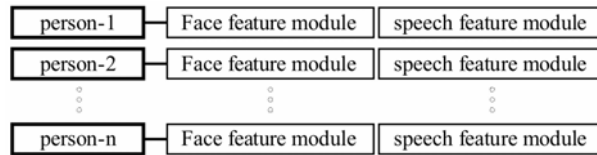


Figure 8. Created personal code book



Figure 9. Some images of registered person

4.3 Proposed fuzzy logic for improved verification

In this subsection, we present a decision method for the face and speech to be certificated using fuzzy logic. The proposed method extracts the candidate lists of recognized face images and speech as shown in Figure 10. In the face, F1 compares three images of the same person with an extracted face candidate. F2 and F3 respectively represent the cases with two and one images. For speech verification, S1 compares three speeches of the same person with an extracted candidate speaker. S2 and S3 respectively represent the cases with two and one speeches. Also, if the extracted candidate of face and speech is same, it is F0&S0 as shown in Figure 10. The verification of face and speech uses Mamdani's fuzzy inference (Manoj et al. 1998).

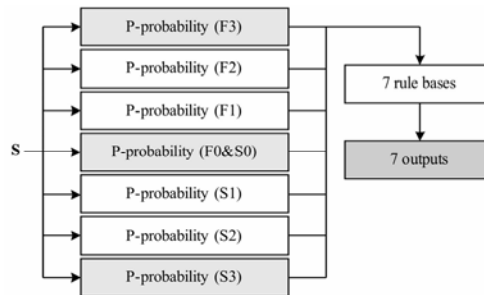


Figure 10. Fuzzy inference engine

The input fuzzy engine contains the recognized probability classified as shown in Figure 10, where $\beta(F3, F2, F1, F0 \& S0, S1, S2, S3)$ represents the coefficient of recognized probability. The basis rule is given as

$$\left\{ \begin{array}{ll} FACE & 1.0 \\ \text{If } P(R) \text{ is COMPLETE} \text{ Then } O_{\theta} \text{ is } 0.5' & \\ SPEECH & 0.0 \end{array} \right. \quad (6)$$

where $R \in \{F3, F2, F1, F0 \& S0, S1, S2, S3\}$, and O_θ represents a pre-specified threshold. The input membership function of fuzzy inference engine is shown in Figure 11. Finally, the predicted human verification result can be stored by using the Singleton's fuzzifier, the product inference engine, and the average defuzzifier as

$$P_{max} = F3(1/O_{F1}) + F2(1/O_{F2}) + \dots + S3(1/O_{S3}). \tag{7}$$

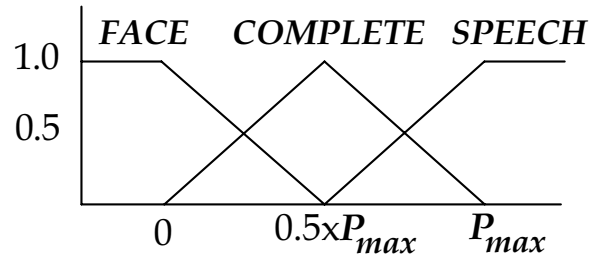
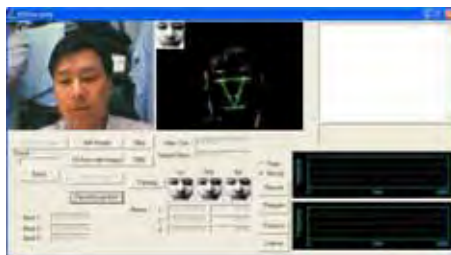


Figure 11. Input membership function of fuzzy engine

5. Experimental results

The proposed multimodal, biometric human recognition system is shown in Figure 12, which shows the result of face and speech extraction. Figure 13 shows the result of registered personal verification. Figure 14 shows the result of non registered person rejection.



(a) face detection and registration



(b) speech detection and registration

Figure 12. The process to recognize face and speech



(a) person verification



(b) person verification

Figure 13. Accepted results



Figure 14. Rejected result for face and speech

The experimental result for the verification rate using the proposed method is summarized in Table 1. An experimental result of FAR given in Table 1 corresponds to 0.01%. In this case, the FAR can accept a person out of 100. Table 2 shows the result of the verification rate and FAR for the proposed method. As shown in Table 2, the proposed method can reduce FAR to 0.0001% and the impersonation to one person out of 10,000. Figure 15 shows that the proposed method can further reduce the equal error rate (EER).

Test DB	verification rates(%)		FAR(%)	
	male	female	male	female
face	98.5		0.01	
speaker	97.37		0.01	

Table 1. Verification rates of male and female

Test DB	verification rate(%)	FAR(%)
face & speaker	99.99	0.0001

Table 2. Verification rate of the proposed method

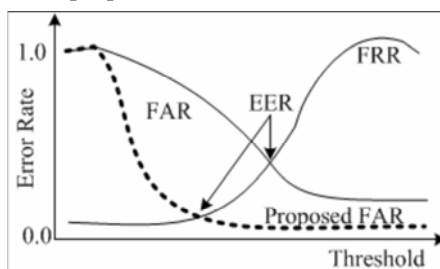


Figure 15. Error rate of the proposed method

6. Conclusions

In this chapter, we present a human verification method using combined face and speech information in order to improve the problem of single biometric verification. Single biometric verification has the fundamental problems of high FAR and FRR. So we present a

multimodal, biometric human verification method to improve the verification rate and reliability in real-time. We use PCA for face recognition and HMM for speech recognition for real-time personal verification. As a result the proposed verification method can provides stable verification rate, and it overcomes the limitation of a single mode system. Based on the experimental results, we show that FRR can be reduced down to 0.0001% in the human multimodal interface method using both face and speech information.

7. References

- Belhumeur, P.; Hespanha, J. and Kriegman, D. (1997) Eigenfaces vs fisherfaces: recognition using class specification linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, (July 1997) Page(s) :711-720, 0162-8828.
- Gu, Y.; Thomas, T. (1999). A hybrid score measurement for HMM-based speaker verification. *Proceedings of IEEE International Conference Acoustics, Speech, and Signal Processing*, Vol. 1. pp. 317-320, March 1999.
- Kim, Y.; Park, C. and Paik, J. (2004). A new 3D active camera system for robust face recognition by correcting pose variation. *Proceedings of International Conference Circuits and Systems*, pp. 1482-1487, August 2004.
- Kong, S.; Heo, J., Abidi, B., Paik, J., and Abidi, M. (2005). Recent advances in visual and infrared face recognition - A review. *Computer Vision and Image Understanding*, Vol. 97, No. 1, (January 2005) Page(s) :103-135, 1077-3142.
- Kriegman, D.; Yang, M. and Ahuja, N. (2002). Detecting faces in images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 1, (January 2002) Page(s) :34-58, 0162-8828.
- Liu, X.; Chen, T. and Kumar, V. (2002). On modeling variations for face authentication. *Proceedings of International Conference Automatic Face and Gesture Recognition*, pp. 369-374, May 2002.
- Manoj, T.; Leena, J. and Soney, R. (1998). Knowledge representation using fuzzy petri nets-revisited. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 10, No. 4, (August 1998) Page(s) :666-667, 1041-4347.
- Osuna, E.; Freund, R. and Girosi, F. (1997). Training support vector machines: an application to face detection. *Proceeding of IEEE Conference Computer Vision and Pattern Recognition*, pp. 130-136, 1997.
- Rabiner, L.; Juang, B. (1998). *Fundamentals of speech recognition*, Prentice-Hall, 1993.
- Rowley, H.; Baluja, S. and Kanade, T. (1998). Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 1, (January 1998) Page(s) :203-208, 0162-8828.
- Samaria, F.; Young, S. (1994). HMM based architecture for face identification. *Image and Vision Computing*, Vol. 12, No. 8, (October 1994) Page(s) :537-543, 0262-8856.
- Sirivich, L.; Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces. *Journal Optical Society of America A: Optics, Image Science, and Vision*, Vol. 4, No. 3, (March 1987) Page(s):519-524.
- Zhang, J.; Yan, Y. and Lades, M. (1997). Face recognition: eigenface, elastic matching, and neural nets. *Proceedings of IEEE*, Vol. 85, No. 9, pp. 1423-1435, September 1997.