# Multimodal Person Verification from Video Sequences

H. K. Ekenel, S. Y. Bilgin, İ. Eden, M. Kirişçi, H. Erdoğan, A. Erçil
Sabancı Üniversitesi, Faculty of Engineering and Natural Sciences
Tuzla, Istanbul 34956 TURKEY
ercil@sabanciuniv.edu

## ABSTRACT

In this paper, a multimodal person verification system based on fusing information derived from face and speech signals is proposed. Principle component analysis and independent component analysis techniques are used for face verification and mel-frequency-cepstral coefficients are used for speaker verification. The matching scores from individual modalities are combined using the sum rule. The results indicate that fusing indivual modalities improve the overall performance of the verification system.

## 1. INTRODUCTION

In the last decade, biometric person verification has found wide range of applications. Biometric systems using a single biometric trait for authentication purposes have some limitations. Noisy data, limited degrees of freedom, spoof attacks, and unacceptable error rates, all affect the performance, security and convenience of using such a system. Multibiometric systems that use multiple traits of an individual for authentication, alleviate some of these problems while improving verification performance.

Despite intensive efforts to solve the face recognition/verification problem since 1970s, the problem still remains unsolved due to large variations in facial appearance caused by, for instance, changes in expression, illumination, occlusion and pose. In initial efforts to build automatic face recognition systems, feature-based methods were popular [11-13]. After 1990s, template matching-based methods have become popular [4]. Since face data is obtained by lexicographically ordering the image pixels, it constitutes a very high dimensional space. To reduce the excessive dimensionality and to capture or approximate the face manifolds, several subspace analysis tools are utilized. The most widely used subspace analysis tools are principle components analysis (PCA) [17, 20, 21, 25], Linear Discriminant Analysis (LDA) [3], Independent Component Analysis (ICA) [1], their nonlinear varieties via kernel tools [16, 23, 28, 29] and their mixture models [8, 14, 15, 24].

Speaker verification based on audio is a relatively mature research area. Similar to face verification, there are problems associated with the speaker verification, such as misspoken or misread phrases, emotional states, sickness and aging. Furthermore, errors during the removal of silence parts from the input speech signal lead performance reductions. In this respect, the features extracted from the speech signal should be robust against these problems.

For speaker verification, many different features such as linear predictive coefficients (LP), cepstrum, and mel-frequency-cepstral coefficient (MFCC) features have been used [5]. Modeling method varies depending on whether one would like to perform text-independent or text-dependent speaker verification. For text-independent verification, nonparametric probability density functions (pdf) and Gaussian mixture models (GMMs) have been used. For text-dependent verification, dynamic time warping (DTW), GMMs and Hidden Markov Models (HMM) are the preferred modeling techniques. Usually, simple Bayes' classification or Neural networks are used for classification purposes [5].

Multimodal biometric systems [9] are expected to be more reliable due to the presence of multiple pieces of evidence. A large number of information fusion methods that can be used for combining evidence from unimodal systems have been proposed in the literature [2, 18, 19, 26, 27].

In this paper, we propose to use the most natural and acceptable biometric modalities for person verification with the emphasis on fusion of the information derived from these sources. Mainly, we will study robust face and speaker verification. For face verification we used PCA and two architectures of ICA. PCA is used as a baseline for comparison purposes. ICA approach has been priorly used for face recognition, but its performance in verification has not been thoroughly tested. Encouraging results in recognition performance have led us to use the ICA approach for face verification. For speech verification, we have used the standart MFCC features and GMM modeling. For combination of the two modalities, we have used the sum rule.

The organization of the paper is as follows: In sections 2 and 3, the unimodal verication techniques are explained. In section 4, the combination scheme used is conveyed. Experimental setup and results are presented in section 5 and conclusions are given in section 6.

## 2. FACE VERIFICATION

### 2.1. Principal Component Analysis

Principal component analysis (PCA) is the most popular subspace projection technique used for face recognition [17, 25]. PCA extracts the linear projection that maximizes the total scatter of the face images. In other words, PCA aims to determine a new orthogonal basis vector set that best reconstructs the face images in the

mean-square error sense. These orthogonal basis vectors, also called eigenfaces, are the eigenvectors of the covariance matrix of the face images, associated with the highest eigenvalues.

### 2.2. Independent Component Analysis

Briefly, ICA is the separation of independent sources from their observed linear mixtures by using high order statistics [10]. In the ICA method, the only information we have is the observations, and neither the mixing matrix nor the distribution of the sources is known. Using the assumptions that the sources are statistically independent and non-Gaussian (at most one of them can have Gaussian distribution), a separation matrix is estimated. Two different architectures are presented for face recognition using ICA [1]. In the first architecture (ICA1), basis images are assumed to be statistically independent whereas in the second architecture (ICA2), the representation coefficients are assumed to be statistically independent. Source images obtained in the first architecture are spatially local and sparse in appearance, while in the second architecture, source images tend to have global face appearance.

## 3. VOICE-BASED VERIFICATION

Features for voice-based verification should characterize the speaker's voice and should distinguish it from other speaker's voices. Short-time spectrum of speech has information about the spoken sounds and speaker's characteristics together. Mel-frequency cepstral coefficients (MFCC) estimate the logarithm of the energy in nonuniformly located frequency bands placed according to the speech perception of humans. After filterbank outputs are obtained, their DCT transform is taken to further decorrelate the feature vector.

$$c_k = \sum_{j=1}^{N} m_j \cos\left(\frac{\pi k}{N}(j-0.5)\right) \qquad (1)$$

Here, $c_k$ are the MFCC coefficients and $m_j$ are the filterbank outputs.

Vocal-tract shape and vocal fold frequency (pitch) are biometrics for a person. Hence, formant locations in the spectrum and pitch for each speaker are person-distinguishing features. MFCC's carry information about the vocal-tract shape but average out the pitch information. Pitch information can be added as another feature as well. However, we use just MFCC features for now and leave adding other features as future work.

For modeling the feature vector from each speaker, a Gaussian mixture model (GMM) is mostly used because of its modeling capability and computational ease. Intuitively, each mixture in a GMM models a different sound in the speaker's speech. There is a well-known expectation-maximization (EM) algorithm for parameter estimation of GMMs. For initialization of GMM means, k-means algorithm is used first to cluster feature vector data. In the end, for each speaker we obtain a joint PDF of the feature vector.

$$f(\mathbf{x} \mid C) = \sum_{k=1}^{K} c_k \, \mathrm{N}(\mathbf{x}, \boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_\mathbf{k}), \qquad (2)$$

where $c_k$ are the mixture weights and $\mathrm{N}(\mathbf{x}, \boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_\mathbf{k})$ are the individual Gaussians for speaker *C*. $\boldsymbol{\Sigma}_\mathbf{k}$ is chosen to be a diagonal matrix for computational reasons.

During testing, we assume a scenario as follows. The user claims an identity *C*. We calculate the frame-averaged log-likelihood of the new feature vectors using the likelihood function given above for the claimed speaker as follows:

$$L_c = \frac{1}{N}\sum_{i=1}^{N} \log f(\mathbf{x}_i|C) = \frac{1}{N}\sum_{i=1}^{N}\left(\log \sum_{k=1}^{K} c_k N(\mathbf{x_i}, \boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_k)\right) \qquad (3)$$

where *N* is the number of feature vectors (frames) extracted from the test speaker and $L_c$ is the log likelihood of the observations belonging to the claimed class.

For decision, the Bayes' optimal solution is to compare likelihood-ratio of hypotheses with a threshold. This requires us to compute the likelihood of the competing hypothesis. We approximate the likelihood of the competing hypothesis, by using a global model obtained by lumping together all feature vectors in the training data.

The frame-averaged log-likelihood from the global model is given as: $L_G = \frac{1}{N}\sum_{i=1}^{N} \log f(\mathbf{x}_i|G)$, (4)

where G denotes the class of all speakers.

The verification system is then carried out by comparing logarithm of the likelihood-ratio $L_c - L_G$ with a threshold. Our current system will serve as a base for our future experiments.

## 4. FUSING FACE AND VOICE MODALITIES

The three possible levels of fusion for multiple biometric traits are: (a) fusion at the feature extraction level -combining features extracted from the raw measurements obtained from each sensor-, (b) fusion at the matching score level -combining partial soft (for instance a continuous score between 0 and 1) decisions, given by the different experts -, (c) fusion at the abstract level -combining hard (accept/reject,or 0/1) decisions of several experts. A majority vote scheme can be used to make the final decision-.

In this work, we decided to utilize the sum rule for matching score level fusion. We will consider that all experts output their local decisions by generating scores in the interval [0,1]. These scores are a measure of their respective belief of the acceptability of the identity claim: the higher the score, the higher the belief that the identity claim is genuine.

An important aspect that has to be considered when combining several experts is the normalization of the scores obtained from these experts (Brunelli and Falavigna,1995). The responses of the different

classifiers usually have different scales (and possibly offsets), so that a sensible combination of the outputs can proceed only after the scores are properly normalized. Normalization typically involves mapping the scores obtained from multiple domains into a common domain before combining them.

A first step towards the normalization of the scores is to reverse the sign of distances, thereby making them concordant with the matching scores. A simple way to normalize scores is to estimate their average values and standard deviations so that they can be scaled into a standard interval, such as [0,1], by means of an appropriate mapping. We have obtained the normalized score $S_{ij}^{''}$ from the original scores $S_{ij}$ ,where i =1,...,d denotes the expert and j denotes a particular score, by using the sigmoid function

$$S_i' = \frac{1}{1+\exp\left(-\frac{S_i-\mu_i}{\sigma_i}\right)} \quad (5)$$

The sum rule has its base in Bayesian theory with the assumption that the posterior probabilities computed by the classifiers do not deviate dramatically from the prior probabilities. The rule can be formalized as follows:

$$\text{Assign } x_n \rightarrow w_c \text{ if } (1-K)P(w_c)+\sum_{k=1}^{K} P(w_c \mid x_{nk}) > \theta \quad (6)$$

where $x_{nk}$ is a specific feature vector, $x_n$ is the sample test pattern, $w_c$ is the claimed class.

Sum rule simply takes the weighted average of the individual score values. Kittler et al. showed in [18] theoretically that under certain assumptions and restrictions many combination schemes often used, such as max, min and average are the special cases of the sum and the product rules. It has also been shown in [18] empirically in two applications that the sum rule is more robust against noise and other disturbances than the product rule, and often outperforms other combination methods. The sum rule has recently been used in [22] with a number of other decision fusion methods, giving the best performance among all. The block diagram of the multi modal biometric system is given in Figure 1.

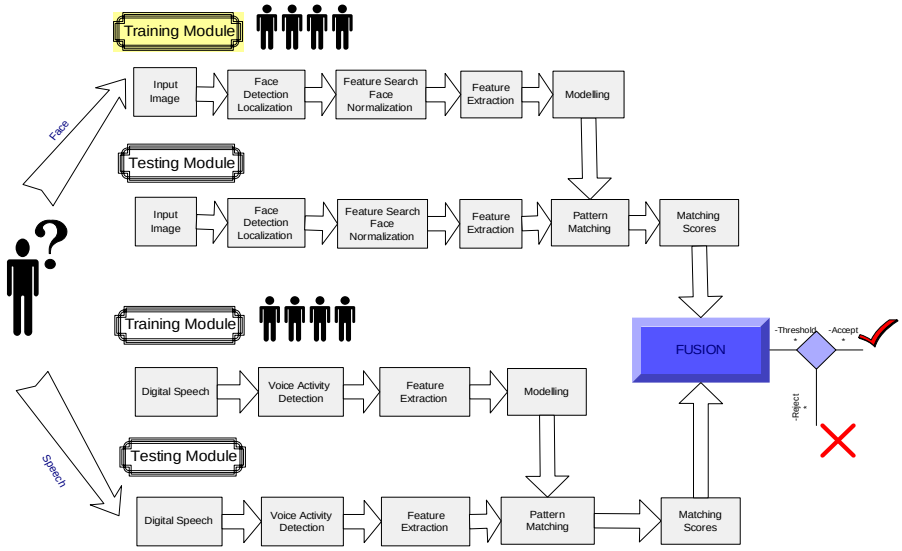


Figure 1. Block diagram of the overall system.

## 5. EXPERIMENTS

### 5.1 Experimental Setup

In our experiments, we used a database of 51 people [30]; 15 video sequences were taken from each subject. Each subject utters 10 repetitions of her/his name and 5 different names from the database. Face images were captured using a Sony SDR-PD150P video camera, with a resolution of 720x576, at a rate of 15 fps and the audio stream has 16 kHz sampling rate. Eye locations of the faces are manually marked and the faces were automatically cropped and aligned. They are then resized to a common resolution of 60x50. Some examples of cropped and resized images can be seen in Figure 2.

Figure 2. Sample cropped and resized images from the database (First two images are from one sequence and the last two images are from a different sequence of the same person)

The videos have been subsampled and only 5 frames/video are used in the experiments. Out of 10 repetitions of the user's name, the first 5 are used for training; the remaining 5 pairs and the impostor data are used for testing.

The thresholds used in the experiments can be person dependent or independent. In the experiments detailed in this paper, a global threshold has been used.

### 5.2 Single biometric experiments

*5.2.1. Face Verification Using Eigenfaces*

When testing the eigenface based face verification, we mainly had two parameters to tune for optimal performance: number of eigenfaces used for representation, and the distance metric. We used three different distance metrics, $L_1$, $L_2$ and normalized cross correlation and we used 20 eigenfaces during our tests.

The test procedure used for face verification from video sequences can be summarized as follows:

1. For each frame in the test video, we calculate the distance between the extracted features and the features of the claimed identity in the training database.
2. We find the minimum distance feature vector.
3. We form a histogram of these distances.
4. Since we have multiple frames for a test session, we also perform fusion at this step. To properly extract the matching scores of each frame, we normalize the distances using a sigmoid function.
5. If the average score is greater than a prescribed value, we authenticate the user.

As can be seen from the distance distribution in Figure 3, the overlap between the genuine and imposter classes is minimal.

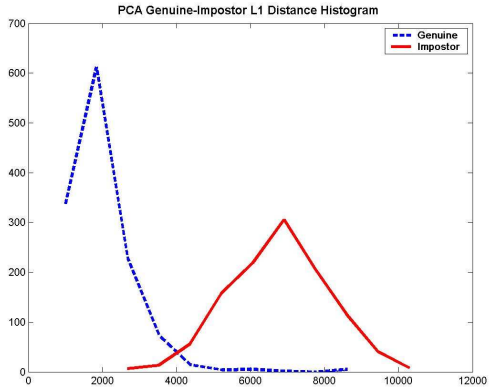


Figure 3. Distance distribution

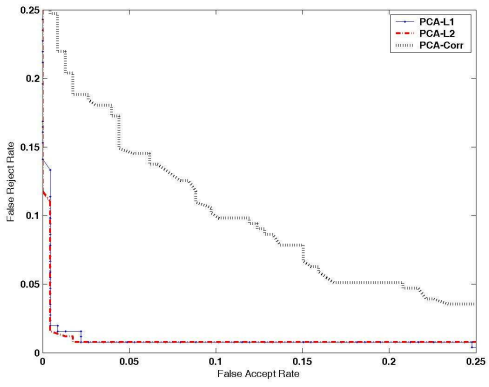


Figure 4. ROC curve for PCA based face verification

It can be observed from Figure 4 that face verification rates using PCA are very satisfactory. (≈ 1% equal error rate, see Table 1). The reason of this performance can be attributed to the fact that proper alignment of the faces is crucial for PCA. $L_1$ and $L_2$ norms perform better than the normalized correlation metric.

Table 1. Equal error rates (EER) for PCA based face verification

| PCA | EER (%) |
|---|---|
| $L_1$ | 1.45 |
| $L_2$ | 1.25 |
| Normalized Corr. | 9.99 |

*5.2.2. Face Verification Using Independent Component Analysis*

As can be observed from the ROC curve in Figure 5, ICA1 performance results are very similar to those of PCA. See Table 2 for equal error rates).

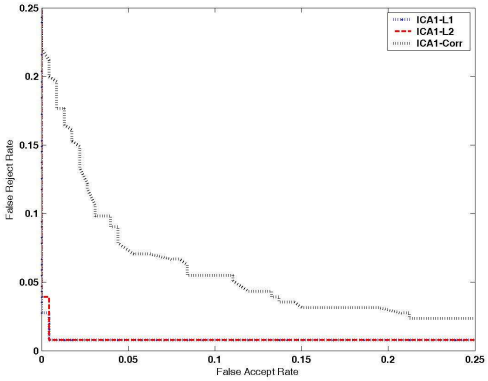


Figure 5. ROC curve for ICA1 based face verification

Table 2. Equal error rates (EER) for ICA1 based face verification

| ICA1 | EER (%) |
|---|---|
| $L_1$ | 0.83 |
| $L_2$ | 0.83 |
| Normalized Corr. | 7.09 |

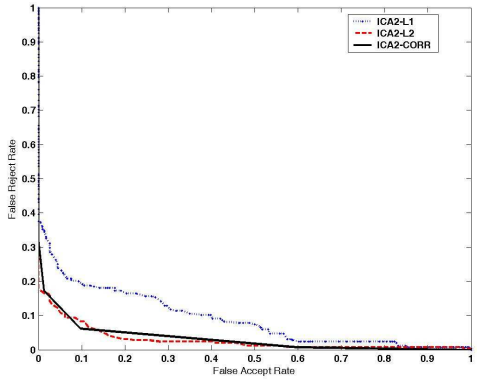


Figure 6. ROC curve for ICA2 based face verification

Note that ICA2 results are inferior to those of PCA and ICA1. This result is surprising considering the fact

that ICA2 performance for face recognition has been shown to be better than PCA and ICA1 [6, 7 ].

Table 3. Equal error rates (EER) for ICA2 based face verification

| ICA2 | EER (%) |
|---|---|
| $L_1$ | 17.87 |
| $L_2$ | 8.96 |
| Normalized Corr. | 8.00 |

*5.2.3. Voice based Verification Experiments*

In this paper, we use standard MFCC based features and Gaussian Mixtures to model each speakers' data regardless of spoken text. In speaker verification applications, it is important to normalize data and perform frame selection to increase performance. We perform a voice activity detection scheme to remove silences from speech data to achieve a simple frame selection. We do not perform any other normalization yet.

We extract an MFCC feature vector every 10 ms from a window of length 25 ms of speech. For training the models for each speaker, we lump together all features extracted from a speaker (across all training data) and train a Gaussian Mixture Model (GMM) with varying number of mixtures in it. We use the well-known EM algorithm for GMM training. We initialize the mixture means using the k-means algorithm. In the end, for each speaker we obtain a joint pdf of the feature vector.

Figure 7 shows the ROC curve obtained using 13 MFCC coefficients modeled with 8 Gaussian mixtures.

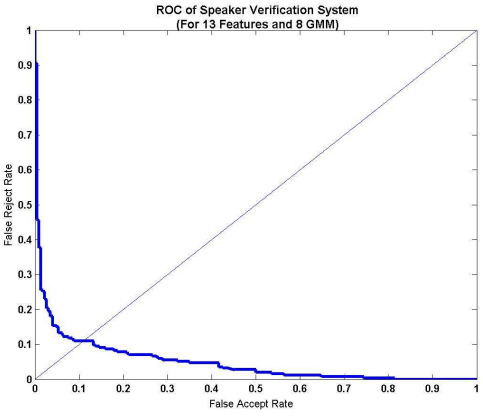


Figure 7. ROC curve for voice based person verification

Table 4. Equal error rates (EER) for speaker verification

| Feature Size | No. of Mixtures | EER (%) |
|---|---|---|
| 13 | 8 | 10.27 |
| 13 | 16 | 10.34 |
| 39 | 8 | 10.74 |
| 39 | 16 | 10.25 |

The equal error rates in Table 4 suggest that there is no significant difference between using 13 MFCC coefficients and its delta coefficients.

**5.3. Fusion Experiments**

In this section, the combination results of the unimodal verification systems using the sum rule are given. We combined the 3 different face verification techniques -taking into consideration only the best performing distance metric– with 13 MFCC coefficients modeled with 8 Gaussian mixtures.

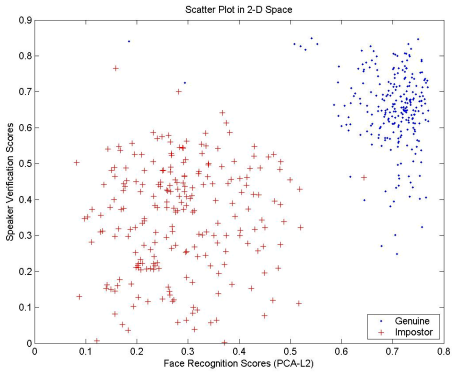


Figure 8. Scatter plot of matching scores using PCA-$L_2$ for face verification and 13 MFCC, 8 mixtures for speaker verification

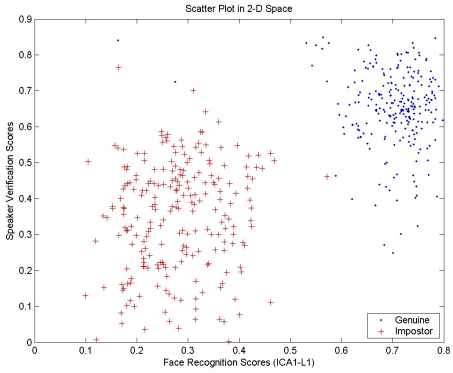


Figure 9. Scatter plot of matching scores using ICA1-$L_1$ for face verification and 13 MFCC, 8 mixtures for speaker verification

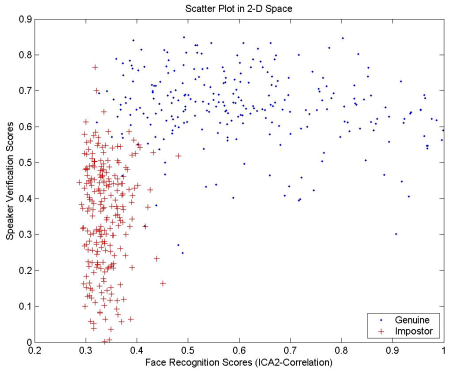


Figure 10. Scatter plot of matching scores using ICA2-Corr for face verification and 13 MFCC, 8 mixtures for speaker verification

It can be seen from Figures 8-10 that the genuine and impostor classes are well separated in the two-dimensional normalized matching scores space. These observations are also supported by the equal error rates depicted in Table 5.

Table 5. Equal error rates (EER) for fusion

| Fusion | EER (%) |
|---|---|
| Speech + PCA $L_2$ | 0.44 |
| Speech + ICA1 $L_1$ | 0.44 |
| Speech + ICA2 Norm. Corr. | 3.10 |

## 6. CONCLUSIONS

In this paper, we have proposed a multimodal person verification system based on fusing information derived from face and speech signals.

For face verification, PCA and ICA1 gave satisfactory results, whereas ICA2 performance was disappointing with respect to its rather successful performance for face recognition tasks.

For speaker verification, 13 MFCC coefficients sufficed for the verification task, and hence there was no need for using delta coefficients.

Combining the two modalities resulted on improved performance rates for all the different combinations studied.

**Acknowledgments**: We would like thank Koc University Multimedia Vision and Graphics Laboratory for providing the dataset used in the paper.

## 7. REFERENCES

[1] M. S. Bartlett, H. M. Lades and T. J. Sejnowski, "Independent Component Representations for Face Recognition", *Proc. of Conf. on Human Vision and Electronic Imaging III*, San Jose, California, 1998.

[2] A. Baykut, "Classifier Combination for Pattern Reconition", Ph.D. Thesis, Bogazici University, 2002.

[3] P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection", *IEEE Trans. on PAMI*, Vol. 19, No. 7, pp. 711-720, 1997.

[4] R. Brunelli, and T. Poggio, "Face Recognition: Features versus Templates", *IEEE Trans. on PAMI*, Vol. 15, No. 10, pp. 1042-1052, October 1993.

[5] J. P. Campbell, "Speaker recognition: A tutorial", *Proc. IEEE*, vol. 85, pp. 1436-1462, Sept 1997.

[6] Draper, B. A., K. Baek, M. S. Bartlett and J. R. Beveridge, "Recognizing Faces with PCA and ICA", *Computer Vision and Image Understanding*, Vol. 91, No. 1-2, pp. 115-137, 2003.

[7] H. K. Ekenel, "Expression and Illumination Insensitive Independent Components and Wavelet Subbands for Face Recognition", M.S. Thesis, Bogazici University, 2003.

[8] B. J. Frey, A. Colmenarez and T. S. Huang, "Mixtures of Local Linear Subspaces for Face Recognition", *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Santa Barbara, *1998.*

[9] L. Hong, A.K. Jain, S. Pankanti, "Can multi-biometrics improve performance?", *Proc. AutoID'99 Summit*, NJ, USA., pp.59 –64, 1999.

[10] A. Hyvärinen, E. Oja, "Independent Component Analysis: Algorithms and Applications", *Neural Networks*, Vol. 13, pp. 411-430, 2000.

[11] T. Kanade,, Picture processing by computer complex and recognition of human faces, Technical Report, Department of Information Science, Kyoto University, Japan, 1973.

[12] Kanade, T., *Computer Recognition of Human Faces*, Birkhauser, Basel and Stuttgart 1977.

[13] M. D. Kelly, "Visual Identification of People by Computer", Technical Report, AI-130, Stanford AI Proj., Stanford, CA, 1970.

[14] H. C. Kim, D. Kim and S. Y. Bang, "Face Recognition Using the Mixture-of-eigenfaces Method**",** *Pattern Recognition Letters*, Vol. 23, No. 13, pp. 1549-1558, 2002.

[15] H. C. Kim, D. Kim and S. Y. Bang, "Face Recognition Using LDA Mixture Model**",** *Pattern Recognition Letters*, Vol. 24, No. 15, pp. 2815-2821, 2003.

[16] K. I. Kim, K. Jung and H. J. Kim, "Face Recognition Using Kernel Principal Component Analysis", *IEEE Signal Processing Letters*, vol. 9, no. 2, pp. 40-42, 2002.

[17] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces", *IEEE Trans. on PAMI*, Vol. 12, No. 1, pp.103-108, January 1990.

[18] J. Kittler, M. Hatef, R.P. Duin, and J.G.Matas, "On combining classifiers", *IEEE Trans. on PAMI*, pp.226-239, 1998.

[19] J. Kittler, "Combining Classifiers: A Theoretical Framework", *Pattern Analysis and Applications*, Vol. 1, No. 1, pp. 18-28, 1998.

[20] B. Moghaddam, T. Jebara and A. Pentland, "Bayesian Face Recognition", *Pattern Recognition*, Vol. 33, No. 11, pp. 1771-1782, November 2000.

[21] A. Pentland, B. Moghaddam, T. Starner and M. Turk, "View based and Modular Eigenspaces for Face Recognition", *Proc. of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 84-91, 1994.

[22] A. Ross, A.K. Jain and J. Qian, "Information fusion in biometrics", *Proc. of Third Intl. Conf. on AVBPA*, Halmstad, Sweden, pp.354-359, 2001.

[23] B. Schölkopf, A. Smola and K.-R. Müller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem", *Neural Computation*, Vol. 10, No. 5, pp. 1299-1319, 1998.

[24] D. S. Turaga and T. Chen, "Face Recognition Using Mixtures of Principal Components", *Proc. of IEEE Intl. Conf. on Image Processing, Rochester*, September 2002.

[25] M. Turk and A. Pentland, "Eigenfaces for Recognition", *Journal of Cognitive Science*, pp. 71-86, 1991.

[26] P. Verlinde, "A contribution to multi-modal identity verification using decision fusion" , PhD. Thesis, Ecole Nationale Superiore de Telecommunications, 1999.

[27] L. Xu, A. Krzyzak and C. Y. Suen, "Methods for Combining Multiple Classifiers and Their Application in Handwritten Character Recognition", *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. 22, pp. 418-435, 1992.

[28] M. H. Yang, N. Ahuja and D. Kriegman, "Face Recognition Using Kernel Eigenfaces", *Proc. of IEEE Intl. Conf. on Image Processing (ICIP 2000)*, Vancouver, September 2000, vol. 1., pp. 37-40, 2000.

[29] M. H. Yang, **"**Kernel Eigenfaces vs. Kernel Fisherfaces: Face Recognition Using Kernel Methods", *Proc. of the Fifth IEEE Intl. Conf. on Automatic Face and Gesture Recognition*, Washinton D.C., May 20 - 21 2002**,** pp. 215-221, 2002.

[30] Y. Yemez, A. Kanak, E. Erzin, and A. M. Tekalp, "Multimodal Speaker Identification With Audio-Video Processing", *Proc. of the Intl. Conf. on Image Processing*, (ICIP 2003), pp. 14-17, September 2003.