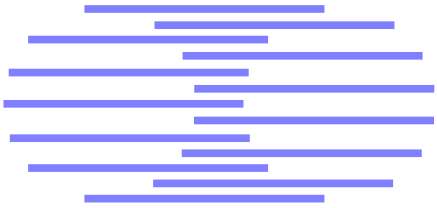
IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse



Fusion of Face and Speech Data for Person Identity Verification

S. Ben-Yacoub [a] Y. Abdeljaoued [b]
E. Mayoraz [b]

IDIAP–RR 99-03

January 1999

submitted for publication

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] IDIAP, CP 592, 1920 Martigny, Switzerland
[b] EPFL-LTS, Ecublens CH-1015 Lausanne, Switzerland

IDIAP Research Report 99-03

Fusion of Face and Speech Data for Person Identity Verification

S. Ben-Yacoub Y. Abdeljaoued E. Mayoraz

January 1999

submitted for publication

Abstract. Multi-modal person identity authentication is gaining more and more attention in the biometrics area. Combining different modalities increases the performance and robustness of identity authentication systems. The authentication problem is a binary classification problem. The fusion of different modalities can be therefore performed by binary classifiers. We propose to evaluate different binary classification schemes (SVM, MLP, C4.5, Fisher's linear discriminant, Bayesian classifier) on a large database (295 subjects) containing audio and video data. The identity authentication is based on two modalities: face and speech.

1 Introduction

The area of identity recognition has been receiving a lot of attention in the last years. There is an increasing demand of reliable automatic user identity recognition systems for secure accesses to buildings or services. Classical techniques based on passwords and cards have a certain number of drawbacks. Passwords may be forgotten or compromised, cards may be lost or stolen and the system is not able to make the difference between a client and the impostor. A lot of techniques have been suggested and investigated by different researchers to recognize users by characteristics which are difficult to impost. Biometrics [12] is the area related to person recognition by means of physiological features (fingerprints, iris, voice, face etc...).

A biometric person recognition system can be used for person identification or verification. In the verification task, a user claims a certain identity (" I am user X"). The system should accept or reject this claim (decide if the user is who he claims to be). In the identification task, there is no identity claim from the user. The system should decide who the user is (eventually unknown in an open-set case). In this work we will focus on the issue of biometric person *verification*.

A large number of commercial biometric systems are using fingerprint, face or voice. Each modality has its advantages and drawbacks (discriminative power, complexity, robustness, etc...). Fingerprint verification has been used for a long time. It is based on local properties of ridges and furrows on the fingertip[20]. The features, called minutiae [13], are extracted and compared to determine possible matches. The image quality of the fingerprints is very important for minutiae extraction. The matching should also cope with problems like cuts on fingertips.

Identification through voice and face is natural and easily accepted by end-users. A lot of work has been done in the last years in the field of face and speaker recognition yielding mature techniques that can be used in applications.

Automated face recognition has been witnessing a lot of activity during the last years [8, 6, 26]. A certain number of new techniques were proposed. Among those, which are representative of new trends in face recognition, one may cite Eigenface [15, 31, 24], elastic graph matching [19], auto-association and back-propagation neural nets [9]. These three techniques were analyzed and evaluated by Zhang et al. [36]. This survey is perhaps the most effective representative and comprehensive because of the analysis of these algorithms under a common statistical decision framework and the evaluation on a common database with more than hundred different subjects. The experimental results of this survey indicate that the *Elastic Graph Matching* (EGM) outperforms other techniques. This method will be presented in Section 2.

Speaker recognition is a very natural way for solving identification and verification problems. With largely available telephone networks and cheap microphones on computers, user recognition through speech becomes a natural solution. A lot of work has been done in this field and generated a certain number of applications of access control for telephone companies [7]. Text-dependent and text-independent speaker verification will be presented in Section 3.

It has been shown that combining different biometric modalities enables to achieve better performances than techniques based on single modalities[5, 11, 16, 14, 2]. Combining different modalities allows to alleviate problems intrinsic to single modalities. The *supervisor* algorithm, which combines the different modalities, is a very critical part of the recognition system. A key question is what strategy should be adopted in order to make the final decision ?

The problem of identity verification is basically a binary classification problem. A potential user claims a certain identity ("I am user X"). The problem of verification consists in deciding whether this claim should be accepted or rejected, and hence is a binary decision problem. The sensed data (face and speech) are processed by different verification experts: a face verification expert and a speaker verification expert. The experts, given the sensed data, will deliver an "opinion" on the user's claim. A final module (the supervisor) will combine the opinions of the different experts and give a binary decision: accept or reject the claim. A classical verification scenario is depicted in Figure 1.

The paper will address the issue of which binary classifier to use in the framework of multi-modal person verification. We propose to investigate different binary classifiers and to evaluate them on a

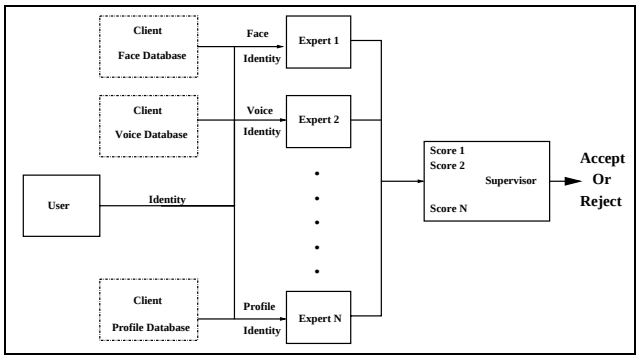


Figure 1: User access scenario

large database (XM2VTS database[1] with 295 people) according to a specified and common testing protocol.

The face verification algorithm will be presented in Section 2. The speaker verification based on text-dependent and text-independent approaches is discussed in Section 3. The fusion of different modalities as well as the different classifiers are described in Section 4. The evaluation protocol and the audio-visual database are presented in Section 5. Finally we present the evaluation results and the main conclusions.

2 Face Verification

The Elastic Graph Matching (EGM) introduces a specific face representation as illustrated in Fig. 1. Each face is represented by a set of feature vectors positioned on nodes of a coarse, rectangular grid placed on the image. As features the modulus of complex Gabor responses from filters with 6 orientations and 3 resolutions are used.

Comparing two faces corresponds to matching and adapting a grid taken from one image to the features of the other image. Therefore both the feature vectors of each node and the deformation information attached to the edges are taken into account. The quality of different matches between an observed grid and a reference grid can be evaluated using the following distance:

$$d(G,R) = \sum_{i=1}^{N_n} d_n(G_{n_i}, R_{n_i}) + \lambda \sum_{j=1}^{N_e} d_e(G_{e_j}, R_{e_j}) \quad (1)$$

$$= \sum_{i=1}^{N_n} d_{n_i} + \lambda \sum_{j=1}^{N_e} d_{e_j} \quad (2)$$

where G_{n_i} represents the ith node of grid G, R_{e_j} is the jth node of grid R; N_n, N_e are the number of nodes and edges, respectively, and λ is a weighting factor which characterizes the stiffness of the graph. A *plastic* graph which opposes no reaction to deformation corresponds to $\lambda = 0$, while a totally rigid graph is obtained with very large values of λ.

Because of the large number of possible matches an approximate solution in [19] was proposed. The matching consists of two consecutive steps: rigid matching and deformable matching. In rigid matching an approximate match is estimated, which corresponds to setting a high value of λ. In deformable matching the grid is deformed in order to minimize (1). Advantages of the elastic graph matching are the robustness against variation in face position, and expression. This owes to the Gabor features, the rigid matching stage, and the deformable matching stage. If the Eigenface is used a scale and face position compensations are needed.

[1] From ACTS-M2VTS project, available at http://www.ee.surrey.ac.uk/Research/VSSP/xm2vts

We note here that the contribution from nodes are considered equally. This is a drawback of the algorithm since the contributions of each node to the distance are different.

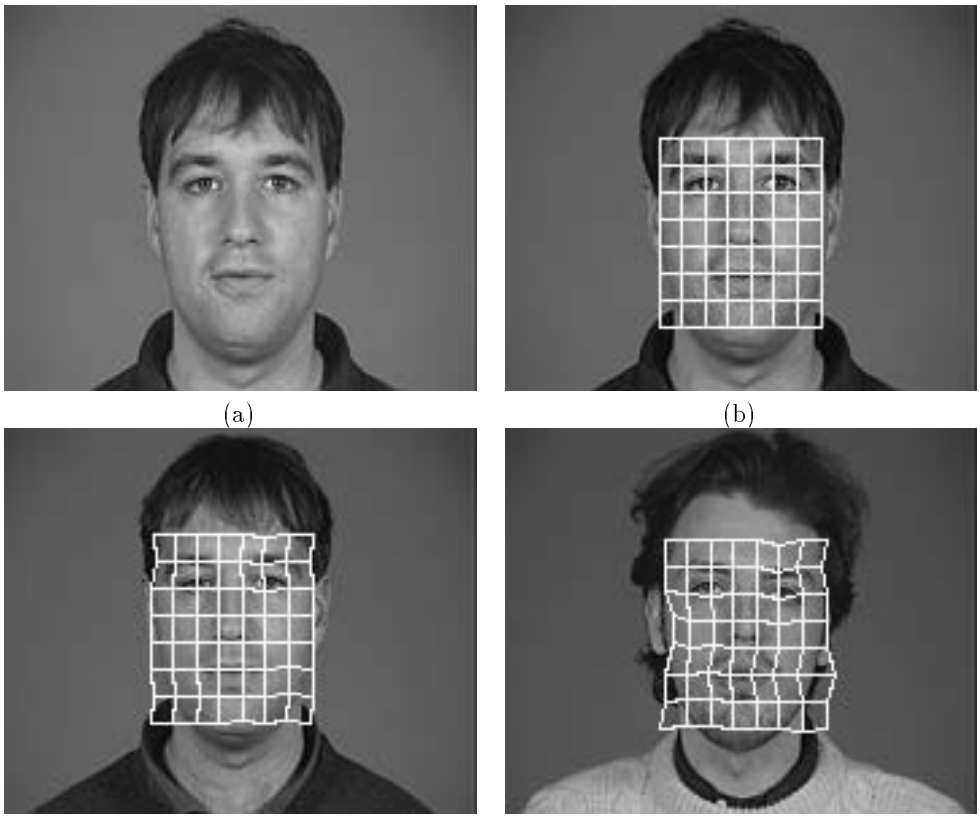


Figure 2: Example of a grid matching. (a) reference image, (b) reference grid, (c) matched grid on another image of the same person, (d) matched grid on another person.

3 Speaker Verification

In the present work, we will use two different approaches to speaker verification. The first is a text-independent based speaker verification approach. The second is based on text-dependent speaker verification: the user has to utter the same text as during the training session (i.e. utterances of digits from 0 to 9). The text-independent method approach uses a sphericity measure [4] and the text-dependent technique uses hidden Markov models (HMM) [29]. Our experiments have used the results of two different speaker verification techniques that have been described in [22].

3.1 Text-independent Speaker Verification

The audio signal (after removal of silence) is converted to linear prediction cepstral coefficients (LPCC) [1]. The energy of the signal is normalized by a mapping to $[0, 1]$ using a tangent hyperbolic function. The feature vector is composed by 12 LPCC coefficients and the signal energy yielding a 13-dimensional vector.

A client is modeled by the covariance matrix $\mathbf{X}$ of the feature vectors of the client's training data

$\{X_1, X_2, ..., X_n\}$:

$$\hat{X} = \frac{1}{n-1}\sum_{i=1}^{n} X_i \quad (3)$$

$$\mathbf{X} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{X})(X_i - \hat{X})^t \quad (4)$$

During a test session, the covariance matrix $\mathbf{Y}$ is computed over the test speech data of a person requesting an access. The arithmetic-harmonic sphericity measure $D_{SPH}(\mathbf{X}, \mathbf{Y})$ [4] is used as similarity measure between the client and the accessing person:

$$D_{SPH}(\mathbf{X}, \mathbf{Y}) = \log\left[\frac{tr(\mathbf{Y}\mathbf{X}^{-1})tr(\mathbf{X}\mathbf{Y}^{-1})}{m^2}\right], \quad (5)$$

where m is the dimension of the feature vector and $tr(\mathbf{X})$ the trace of $\mathbf{X}$. The similarity values were mapped to the interval $[0, 1]$ with a sigmoid function. This similarity measure will be used by the "supervisor algorithm" in order to take the final decision about the person's claim.

3.2 Text-dependent Speaker Verification

The text dependent speaker verification is based on Hidden Markov models (HMMs). The HMMs were largely used in speech processing because of the temporal structure of the speech signal[28]. The HMM has a certain number of parameters that are set so as to best explain a given set of patterns of a known category. We will define two categories: the client category and the impostor (or world) category. Each client's training set will generate a particular HMM (i.e a HMM with a certain instance of its parameters). The world or impostor training set will also generate a particular HMM.

The feature vector that will be used is the same as for sphericity. Temporal informations as first and second derivatives will be added to the feature vector yielding a 42-dimensional vector.

The HMM model of a particular category allows to compute the likelihood of a test pattern or feature vector (i.e given a test pattern what is the likelihood that it was generated by this model). When a user claims a certain identity **Id**, the HMM of the claimed identity will be used to compute the likelihood of the feature vector being generated by the client **Id**. Similarly, the HMM modeling the world (or impostors) will be used to compute the likelihood of the feature vector being generated by an impostor. The decision is then made by comparing the likelihood ratio to a predefined threshold.

The HMM-based verification technique that is used here needs 3 HMM sets: client models, world models, and silence models. The world models serve as speaker-independent models to represent speech of an average person. The world models are computed on a distinct database POLYCOST[2] database. Finally, three silence HMMs are used to model the silent parts of the signal.

All models were trained based on the maximum likelihood criterion using the Baum-Welch (EM) algorithm. The world models were trained on the segmented words of the POLYCOST database, where one HMM per word was trained.

For verification, the Viterbi algorithm [28] is used to calculate the likelihood $p(X_j|\mathcal{M}_{ij})$, where X_j represents the observation of the segmented word j; $\mathcal{M}_{ij}$ represents the model of subject i and word j. The log-likelihood of word j is normalized by the numbers of frames N_j and sum them over all words W, which leads to the following measure:

$$\log p(X|\mathcal{M}_i) = \frac{1}{W}\sum_{j=1}^{W}\frac{\log p(X_j|\mathcal{M}_{ij})}{N_j} \quad (6)$$

[2] For more informations see http://circwww.epfl.ch/polycost

This measure is calculated for the models $\mathcal{M}_c$ of a given client c and for the world models $\mathcal{M}_w$. The following similarity:

$$D_{HMM} = \log \frac{p(X|\mathcal{M}_c)}{p(X|\mathcal{M}_w)} \tag{7}$$

is computed and mapped to the interval $[0, 1]$ as described in Section 3.1. The final measure will be then used by the "supervisor" to make the final decision.

4 Fusion

Having computed a match score between the claimed identity and the user, a verification decision is made whether to accept or reject the claim.

Combining different modalities results in a system which can outperform single modalities [18, 17]. This is especially true if the different experts are not correlated. We expect from the fusion of vision and speech to achieve better results. In the next section, we will investigate different fusion schemes and compare them. The different binary classification approaches that will be evaluated are:

- Support Vector Machines.
- Minimum cost Bayesian classifier.
- Fisher's linear discriminant.
- C4.5 decision trees.
- Multi Layer Perceptron.

4.1 SVM fusion

The Support Vector Machine is based on the principle of *Structural Risk Minimization* [32]. Classical learning approaches are designed to minimize the empirical risk (i.e error on a training set) and therefore follow the *Empirical Risk Minimization* principle. The SRM principle states that better generalization capabilities are achieved through a minimization of the bound on the generalization error.

We assume that we have a data set $\mathcal{D}$ of M points in a n dimensional space belonging to two different classes $+\mathbf{1}$ and $-\mathbf{1}$:

$$\mathcal{D} = \{(\mathbf{x}_k, y_k)|k \in \{1..M\}, \mathbf{x}_k \in \mathbb{R}^n, y_k \in \{+1, -1\}\}$$

A binary classifier should find a function f that maps the points from their data space to their label space.

$$\begin{array}{rccc} f : \mathbb{R}^n & \longrightarrow & \{+1, -1\} \\ \mathbf{x}_k & \longmapsto & y_k \end{array}$$

It has been shown [32] that the optimal separating surface is expressed as:

$$f(\mathbf{x}) \quad = \quad sign(\sum_i \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b) \tag{8}$$

where $K(\mathbf{x}, \mathbf{y})$ is a positive definite symmetric function, b is a bias estimated on the training set, α_i are the solutions of the following Quadratic Programming (QP) problem:

$$\begin{cases} \min_{\mathcal{A}} W(\mathcal{A}) = -\mathcal{A}^t I + \frac{1}{2}\mathcal{A}^t D \mathcal{A} \\ \text{with the constraints:} \\ \sum_i \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0 \\ \\ \text{where:} \\ (i,j) \in [1..M] \times [1..M] \\ (\mathcal{A})_i = \alpha_i \\ (I)_i = 1 \\ (D)_{ij} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \end{cases}$$

The kernel functions $K(\mathbf{x}, \mathbf{y})$ define the nature of the decision surface that will separate the data. They satisfy some constraints in order to be applicable (Mercer's conditions, see [32]). Some possible kernel functions have been already identified (we assume $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^n$) :

- $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^t \mathbf{y} + 1)^d$ with $d \in \mathbb{N}$, this defines a polynomial decision surface of degree d.
- $K(\mathbf{x}, \mathbf{y}) = e^{-g||\mathbf{x}-\mathbf{y}||^2}$ is equivalent to one RBF classifier.

The computational complexity of the SVM during the training depends on the number of data points rather than on their dimensionality. The number of computation steps is $O(M^3)$ where M is the number of data points. At run time the classification step of SVM is a simple weighted sum. The classification of 112400 claims requires 5.6sec on an Ultra-Sparc 30.

4.2 Minimum Cost Bayesian Classifier

Since data from multiple sensors is used for the detection of the identity of a person (signal of interest), we can use results from the fields *distributed detection* and *distributed estimation* [34]. Doing so the problem of the multi-modal person authentication can be formulated using the Bayesian risk [33].

Let us consider the binary event ω, which denotes the presence of the claimed identity ($\omega = 1$) or its absence ($\omega = 0$). Given the a priori probability $g = P(\omega = 1)$, the joint density of the local authentication probabilities obeys

$$\xi(\mathbf{x}) = f(\mathbf{x}|\omega = 0)(1 - g) + f(\mathbf{x}|\omega = 1)g, \tag{9}$$

where $f(\mathbf{x}|\omega)$ is a likelihood function.

Using the Bayes' theorem, the a posteriori authentication probability is

$$p = P(\omega = 1|\mathbf{x}) = \frac{f(\mathbf{x}|\omega = 1)g}{\xi(\mathbf{x})}. \tag{10}$$

By combining (9) and (10) we obtain

$$p = \left\{ 1 + \left[\frac{g}{1-g} \frac{f(\mathbf{x}|\omega = 1)}{f(\mathbf{x}|\omega = 0)} \right]^{-1} \right\}^{-1}. \tag{11}$$

Furthermore assuming that the sensors outputs are independent (the measurements are physically independent),

$$f(\mathbf{x}|\omega) = \prod_{i=1}^{n} f_i(x_i|\omega), \tag{12}$$

where $f_i(x_i|\omega)$ is a local likelihood function of the sensor i.

Using (12) we can express the a posteriori authentication probability p as a function of local likelihood ratios given by:

$$p = \left\{ 1 + \left[\frac{g}{1-g} \prod_{i=1}^{n} \frac{f_i(x_i|1)}{f_i(x_i|0)} \right]^{-1} \right\}^{-1} . \tag{13}$$

The Bayesian formulation of the person authentication problem requires the definition of a function which assigns a cost to each correct and incorrect decision. Specifically, C_{ij}, with $i, j \in \{0, 1\}$, represents the cost of deciding $a = i$ when $\omega = j$ is present. The aim of this method is to minimize the expected cost function, also called the *Bayes risk* [23].

$$\begin{aligned} B &= E\{C_{ij}\} \\ &= C_{00}P(a=0,\omega=0) + C_{01}P(a=0,\omega=1) + \\ &\quad C_{10}P(a=1,\omega=0) + C_{11}P(a=1,\omega=1). \end{aligned} \tag{14}$$

The solution of this optimization problem is given in [23], where the observation vector $\mathbf{y} = (y_1, \cdots, y_n)$ is used instead of the authentication probability vector $\mathbf{x}$.

$$a = \begin{cases} 1 & \text{if } \frac{f(\mathbf{x}|\omega=1)}{f(\mathbf{x}|\omega=0)} > \frac{1-g}{g} \frac{C_{10}-C_{00}}{C_{01}-C_{11}} \\ \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

Since the sensors are independent (15) can be expressed in the following form:

$$a = \begin{cases} 1 & \text{if } \prod_{i=1}^{n} \frac{f_i(x_i|\omega=1)}{f_i(x_i|\omega=0)} > \frac{1-g}{g} \frac{C_{10}-C_{00}}{C_{01}-C_{11}} \\ \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

We note in (16) that the optimal solution for the decision rule is also a likelihood ratio test.

The standard "0-1"cost function [3] has been chosen in this work. The cost is 0 if a correct decision is made, and 1 if an incorrect decision is made. This choice arises from the fact that the decision threshold is easy to determine. Furthermore we assume that $g = 1/2$ (i.e. both events $\omega = 1$ and $\omega = 0$ are equally likely a-priori). In this case the decision rule becomes:

$$a = \begin{cases} 1 & \text{if } \prod_{i=1}^{n} \frac{f_i(x_i|\omega=1)}{f_i(x_i|\omega=0)} > 1 \\ \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

The quality of the probability fusion and decision models depends on the modeling of the likelihood function. Due to its shape diversity and to the domain of its densities $[0, 1]$, the Beta family of distributions is a good candidate for our modeling purposes.

$$\begin{aligned} f_i(x_i|\omega) &= Be(\alpha_{i,\omega}, \beta_{i,\omega}) \\ &= \frac{\Gamma(\alpha_{i,\omega} + \beta_{i,\omega})}{\Gamma(\alpha_{i,\omega})\Gamma(\beta_{i,\omega})} x^{\alpha_{i,\omega}-1}(1-x)^{\beta_{i,\omega}-1}, \end{aligned} \tag{18}$$

where Γ is the gamma function, $0 \leq x_i \leq 1$, $\alpha_{i,\omega} > 0$ and $\beta_{i,\omega} > 0$. The mean μ and the variance σ of the Beta distribution are given by [3]:

$$\mu = \frac{\alpha_{i,\omega}}{\alpha_{i,\omega} + \beta_{i,\omega}} \tag{19}$$

$$\sigma = \frac{\alpha_{i,\omega}\beta_{i,\omega}}{(\alpha_{i,\omega} + \beta_{i,\omega})^2(\alpha_{i,\omega} + \beta_{i,\omega} + 1)}. \tag{20}$$

Once μ and σ are estimated from a training set, the parameters $\alpha_{i,\omega}$ and $\beta_{i,\omega}$ of the likelihood function $f_i(x_i|\omega)$ can be determined by using the relationships (19) and (20).

Since the likelihood function is specified by (18), we can substitute the likelihood ratio in (13) and (17) with:

$$\frac{f_i(x_i|\omega=1)}{f_i(x_i|\omega=0)} = \frac{\Gamma(\alpha_{i,1}+\beta_{i,1})}{\Gamma(\alpha_{i,1})\Gamma(\beta_{i,1})}\frac{\Gamma(\alpha_{i,0})\Gamma(\beta_{i,0})}{\Gamma(\alpha_{i,0}+\beta_{i,0})} \cdot x_i^{\alpha_{i,1}-\alpha_{i,0}}(1-x_i)^{\beta_{i,1}-\beta_{i,0}}$$

4.3 Fischer Linear Discriminant

There is a group of classifiers called linear discriminant classifiers. The main idea of these classifiers is to project n-dimensional data onto a line according to a given direction $\mathbf{w}$. If the direction is chosen correctly, then the classification task can be easier in one dimension. The choice of the projection direction can be determined by different criteria. The Fischer's linear discriminant [10] aims at maximizing the ratio of between-class scatter to within-class scatter.

Given a set of n_1 points belonging to class $\mathcal{C}_1$, and n_2 points belonging to class $\mathcal{C}_2$. We suppose the data points $\mathbf{x}_i$ to be in $\mathbb{R}^n$. The d-dimensional mean of each class is:

$$\mathbf{m}_1 = \frac{1}{n_1}\sum_{k\in\mathcal{C}_1}\mathbf{x}_i$$

$$\mathbf{m}_2 = \frac{1}{n_2}\sum_{k\in\mathcal{C}_2}\mathbf{x}_i$$

The scatter matrices of each class is defined by:

$$\mathbf{S}_1 = \sum_{k\in\mathcal{C}_1}(\mathbf{x}_k-\mathbf{m}_1)(\mathbf{x}_k-\mathbf{m}_1)^t$$

$$\mathbf{S}_2 = \sum_{k\in\mathcal{C}_2}(\mathbf{x}_k-\mathbf{m}_2)(\mathbf{x}_k-\mathbf{m}_2)^t$$

The between-class scatter matrix is defined by:

$$\mathbf{S}_b = (\mathbf{m}_2-\mathbf{m}_1)(\mathbf{m}_2-\mathbf{m}_1)^t$$

The linear discrimination consists in finding a direction vector $\mathbf{w}$ in $\mathbb{R}^n$ for the linear projection. All the data points will be projected as follows:

$$y = \mathbf{w}^t\mathbf{x}$$

The projection direction which guarantees the best separation is given by Fischer's criteria (i.e. maximizing the ratio of between-class scatter to within-class scatter). This criteria can be written as finding $\mathbf{w}$ which maximizes the functional $J(\mathbf{w})$:

$$J(\mathbf{w}) = \frac{\mathbf{w}^t\mathbf{S}_b\mathbf{w}}{\mathbf{w}^t(\mathbf{S}_1+\mathbf{S}_2)\mathbf{w}}$$

The functional is also known as the Rayleigh quotient, and the maximum is reached at :

$$\mathbf{w} = (\mathbf{S}_1+\mathbf{S}_2)^{-1}(\mathbf{m}_1-\mathbf{m}_2)$$

4.4 C4.5 classifier

A decision tree, is a tree where at each node a test on a particular attribute of the data is performed, and where the leafs corresponds to a particular class. The path from the root node to a particular leaf is then a series of tests on the attributes that classifies the data to the class defined by the particular leaf.

C4.5 is the most used algorithm for inducing decision trees [27]. It uses approaches from information theory to derive the most discriminant features. During training, an entropy criteria selects the most informative or discriminative features. The input space is then partitioned recursively. A tree pruning is also performed, it reduces the effect of noise and discards non significant sub-trees. The rules applied during pruning generate a more general description of the classification process.

4.5 MLP classifier

A multi-layer perceptron with one hidden layer will be used for the classification purpose. The hidden layer will be composed by 10 hidden units. Training will be performed with the classical back-propagation algorithm [35, 30].

5 Experiments and Results

5.1 The XM2VTS database

The XM2VTSDB [25] database contains synchronized image and speech data as well as sequences with views of rotating heads. The database includes four recordings of 295 subjects taken at one month intervals. On each visit (session) two recordings were made: a speech shot and head rotation shot. The speech shot consisted of frontal face recording of each subject during the dialogue.

The database was acquired using a Sony VX1000E digital cam-corder and DHR1000UX digital VCR. Video is captured at a color sampling resolution of 4:2:0 and 16bit audio at a frequency of 32kHz. The video data is compressed at a fixed ratio of 5:1 in the proprietary DV format. In total the database contains approximately 4 TBytes (4000 Gbytes) of data.

When capturing the database the camera settings were kept constant across all four sessions. The head was illuminated from both left and right sides with diffusion gel sheets being used to keep this illumination as uniform as possible. A blue background was used to allow the head to be easily segmented out using a technique such as chromakey. A high-quality clip-on microphone was used to record the speech. The speech sequence consisted in uttered digits from 0 to 9.

5.2 The experiments protocol

The database was divided into three sets: training set, evaluation set, and test set (see Fig. 3). The training set is used to build client models. The evaluation set is selected to produce client and impostor access scores which are used to estimate parameters (i.e. thresholds). The estimated threshold is then used on the test set. The evaluation set is used by the fusion module as training set. The test set is selected to simulate real authentication tests. The three sets can also be classified with respect to subject identities into client set, impostor evaluation set, and impostor test set. For this description, each subject appears only in one set. This ensures realistic evaluation of imposter claims whose identity is unknown to the system. Two different configuration are proposed. The main difference is in the choice of sessions for the evaluation set.

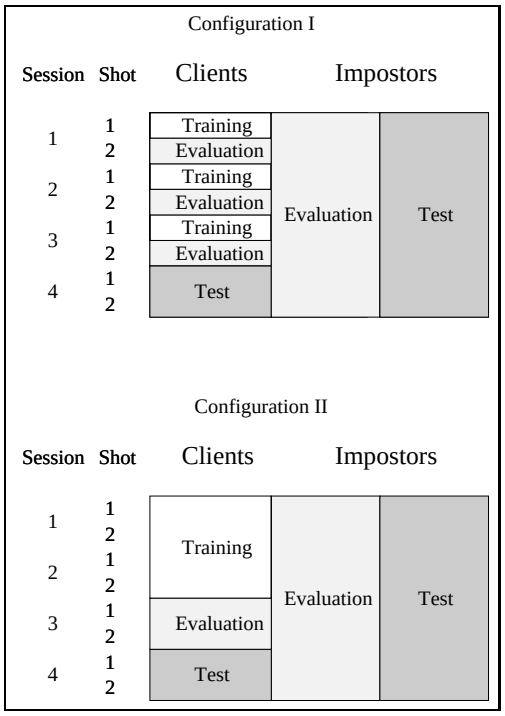


Figure 3: Diagram showing the partitioning of the XM2VTSDB according to protocol Configuration I and II

The protocol is based on 295 subjects, 4 recording sessions, and two shots (repetitions) per recording sessions. The database was randomly divided into 200 clients, 25 evaluation impostors, and 70 test impostors (See [21] for the subjects' IDs of the three groups).

5.3 Performance Measures

Two error measures of a verification system are the *False Acceptance rate* (FA) and the *False Rejection rate* (FR). False acceptance is the case where an impostor, claiming the identity of a client, is accepted. False rejection is the case where a client, claiming his true identity, is rejected. FA and FR are given by $FA = EI/I * 100\%$ and $FR = EC/C * 100\%$, where EI is the number of impostor acceptances, I the number of impostor claims, EC the number of client rejections, and C the number of client claims. FA and FR are functions of a threshold that can control the trade-off between the two error rates. For the protocol configurations, I is $112,000$ (70 impostors $\times$ 8 shots $\times$ 200 clients) and C is 400 (200 clients $\times$ 2 shots).

The performance of the verification system can be also represented by the ROC (receiver operating characteristic), which plots probability of FA versus probability of FR for different values of the threshold. The point on the ROC defined by FA=FR is the *Equal Error Rate* point. Better systems have a ROC curve which is closer to the origin (low FA and FR). The a-priori performance of each modality on the test sets is displayed in Table 1.

The EER of the face verification algorithm is around 8 % for configuration I and 7% for configuration II. The text-independent speaker verification achieves a FA of 1.6% and a FR of 5.0% for configuration I. The EER is around 4% for configuration II. The data for text-dependent speaker verification are available for configuration I only. The error rates on the configuration I are a FA of 0 % and a FR of 1.48%. The fusion results will be compared to the performance of the single modalities.

	Conf I		Conf II	
Modality	FA (%)	FR (%)	FA (%)	FR (%)
Face	8.08	8.5	7.67	7.25
Voice (Spher.)	1.6	5.00	5.53	4.25
Voice (HMM)	0.00	1.48	-	-

Table 1: Performance of Single Modalities on Test Sets (source [22])

We expect from fusion to improve the performance of the whole system. The main motivation of combining different modalities is the increase of performance. This objective can be illustrated by comparing the ROC curves of single modalities to the ROC curve of the combined modalities after fusion. The result is depicted in Figure 4 for the case of Bayesian fusion presented in Section 4.2. The ROC curve of the Bayesian fusion (face,text-dependent speech and text-independent) is clearly outperforming the single modalities.

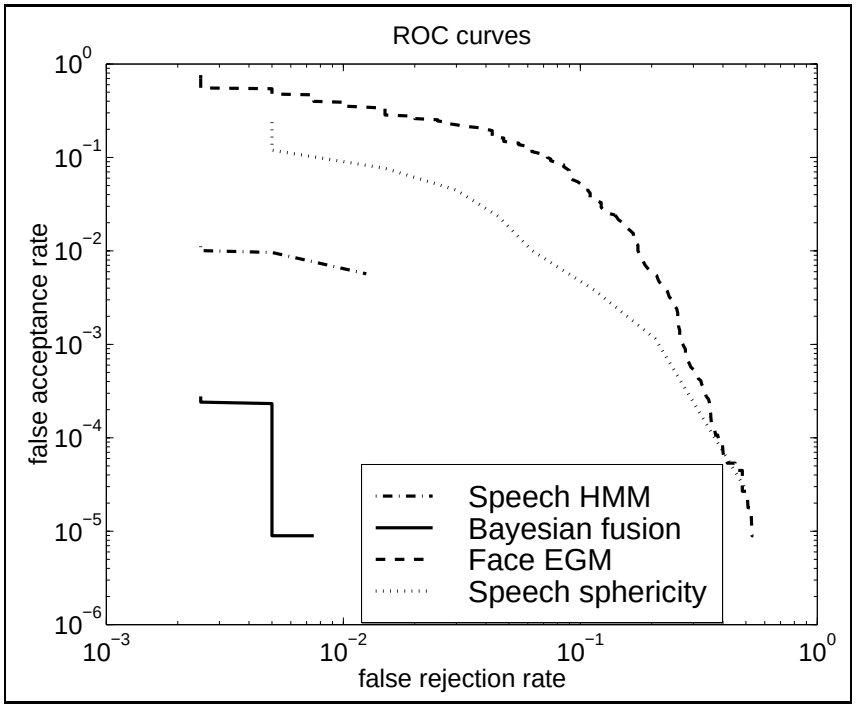


Figure 4: ROC curves of Bayesian Fusion and Single Modalities

The evaluation will be performed on certain combinations of modalities. The sets are defined as follows:

- C1: Face and text-dependent (HMM) in Configuration I.
- C2: Face, text-independent (sphericity) and HMM in Configuration I.
- C3: Face and text-independent (sphericity) in Configuration II.

For the SVM-based fusion, we used polynomial and Gaussian kernels. The training set was used as an evaluation set to see how performance changes with different kernel parameters. For the polynomial kernel set , the degree 2 kernel outperformed the others for the set C1 and C2. The polynomial kernel of degree 5 was the best for set C3.

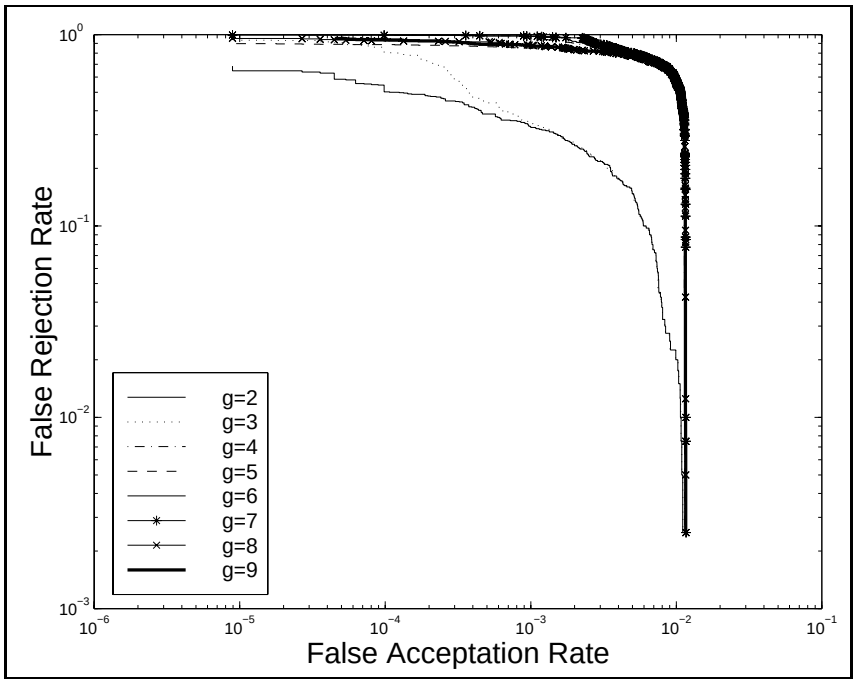


Figure 5: ROC curves of Gaussian Kernels on Evaluation Set of C1

In the Gaussian kernel set, defined by $K(x,y) = exp(-g||x-y||^2)$, the best performance was achieved on set C1 with g=1, on set C2 with g=3 and g=9 for set C3. Figure 5 illustrates the different performances of the Gaussian kernels on evaluation set of C1. The ROC curve clearly shows that the kernel with parameter g=2 outperforms the others.These parameters will be used when comparing the SVM-fusion scheme to the other classifiers.

5.4 Experiments Results

The comparison of the classifiers for the set C1 (Face and text-independent speaker verification) shows that MLP has a very poor performance. The SVM-Gaussian kernel is also not achieving a good performance. The SVM-polynomial kernel and the Fisher linear discriminant generate ROC curves that are very close. For very low FA rates the SVM is behaving slightly better than the Fisher linear discriminant. The Bayesian classifier achieved very good results with a minimum total error rate (i.e. FA+FR) of 0.8% whereas the SVM-polynomial achieved a minimum total error rate of FA of 1.06%.

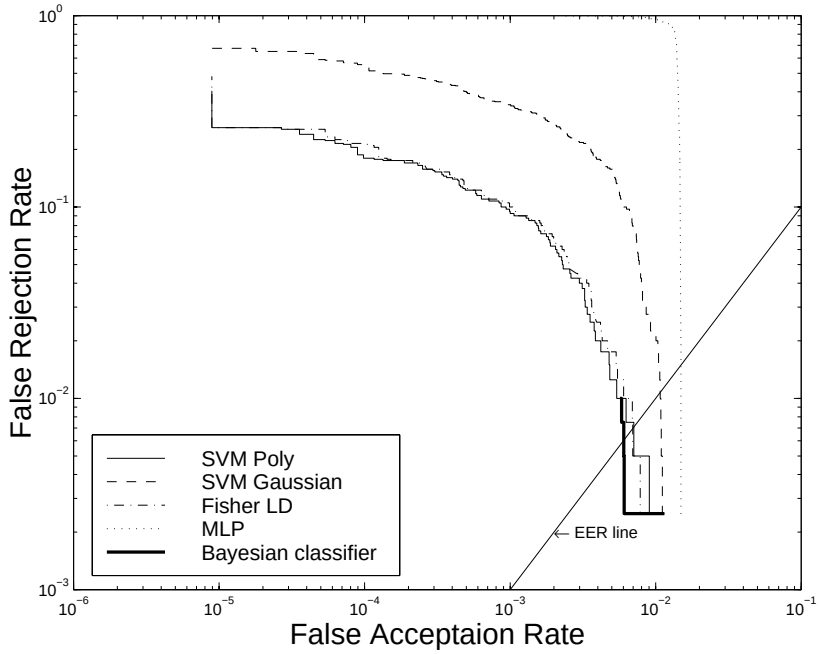


Figure 6: ROC curves for C1

The set C2 combines three modalities (Face, text-dependent and text-independent speaker verification) in configuration I. Here again the MLP does not achieve a good performance. The SVM-Gaussian kernel is even worse than the Fisher linear discriminant. The minimum total error rate for the Fisher classifier is 1.02% and the EER point is 0.68%. The minimum total error rate of the Bayesian classifier is 0.6% and 0.9% for the SVM-polynomial kernel. Both methods achieves a EER point of 0.5%.

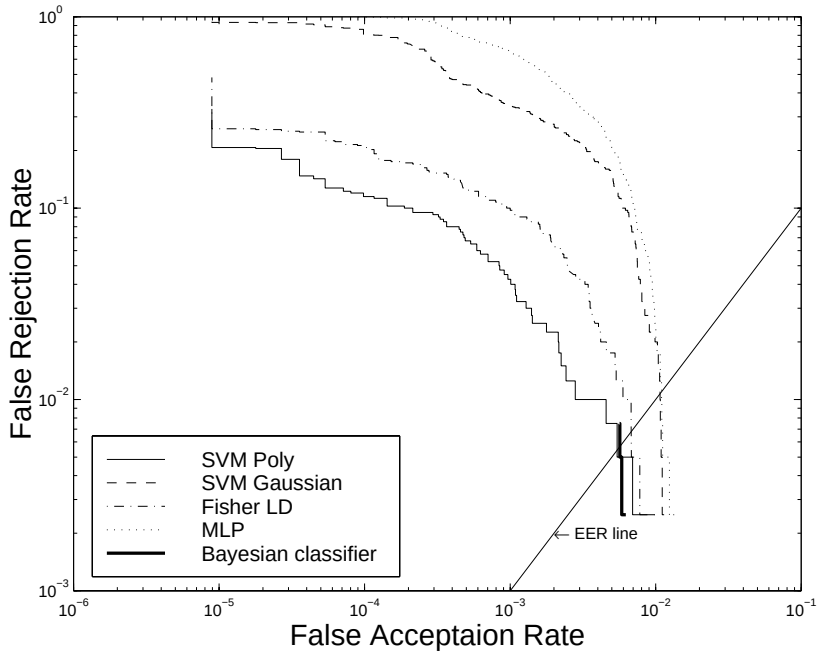


Figure 7: ROC curves for (C2)

	C1		C2		C3	
Classifier	FA (%)	FR (%)	FA (%)	FR (%)	FA (%)	FR (%)
SVM Poly.	1.07	0.25	1.09	0.0	2.09	1.5
SVM Gauss.	1.16	0.0	1.18	0.0	3.26	1.0
Bayesian Fusion	1.21	0.0	0.63	0.0	1.5	1.75
C4.5	0	3.41	0	3.41	0.12	4.75
Fisher LD	1.47	0.0	1.45	0.0	69.2	0
MLP	1.6	0.0	1.58	0.0	0.17	4.0

Table 2: *A-priori* performance of classifiers on test sets

The set C3 is a combination of two modalities (face and text-independent speaker verification) in configuration II. The ROC curves for this set, see Figure 8, are very close and forming a compact group. The Fisher classifier failed in this set and did not provide good results (its ROC curve is not plotted). All the classifiers have almost the same performance with an EER of 1.9%. For low values of FA (less than 1%), the MLP classifier has the lowest FR rate.

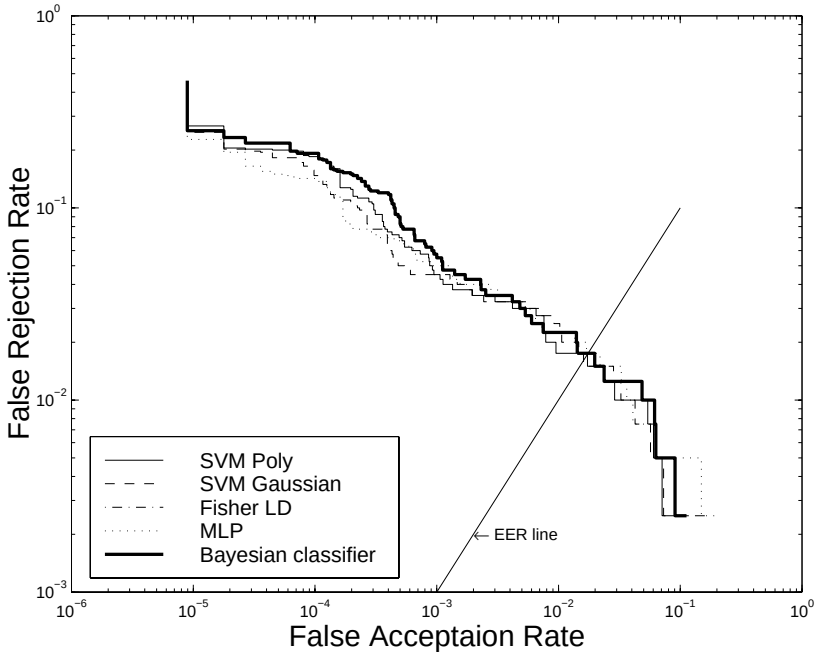


Figure 8: ROC curves for C3

The ROC curves give the result *a-posteriori*, the threshold controlling the trade-off between FA and FR is scanned over the interval of possible values and the corresponding ROC curve is then plotted. Another performance measure consists in fixing the threshold *a-priori* on an evaluation set, and then testing with this value of the threshold on a test set. We compared also *a-priori* results of the different classifiers. The results are displayed on Table 2. This also enables to compare with non-metric classifiers like C4.5 (there is no threshold for controlling FA and FR in C4.5 trees).

6 Conclusion

Multi-modal person verification is a very promising technique. It combines the advantages of different techniques and performs better than single modalities. We described a multi-modal system using face information and speech for user verification. A critical question is how to combine the different modalities. The verification task is a binary classification problem (accept or reject the user's identity claim). We have evaluated different binary classifiers for the fusion of multi-modal data. In order to have a fair evaluation of the different approaches, we compared the performances of the different fusion schemes on a large database (295 subjects) with a specified testing protocol. The training sets for the fusion task consisted in 600 client claims and 40000 impostor claims. The test set (completely independent from the training set) consisted in 400 client claims and 112000 impostor claims.

Among all the classifiers that were evaluated (SVM-Polynomial, SVM-Gaussian, C4.5, MLP, Fisher linear discriminant, Bayesian classifier), the SVM-Polynomial and the Bayesian classifiers showed the best results with a slightly better performance for the Bayesian classifier. The performance of the multi-modal system was considerably increased when compared to the performance of the single modalities.

Acknowledgment

This work was done in the framework of the ACTS-M2VTS european project. The authors also acknowledge the support of the Swiss Federal Office for Education and Science. The authors thank J. Lüttin, G. Maître and D. Genoud for helpful discussions and contributions to the speaker verification algorithms.

References

[1] B.S. Atal. Effectivness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *JASA*, 55(6):1304–1312, 1974.

[2] S. Ben-Yacoub. Multi-Modal Data Fusion for Person Authentication using SVM. In *2nd Intl. Conference on Audio-Video based Biometric Person Authentication*, 1999.

[3] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 1985.

[4] F. Bimbot, I. Magrin-Chagnolleau, and L. Mathan. Second-order statistical measure for text-independent speaker identification. *Speech Communication*, 17(1-2):177–192, 1995.

[5] R. Brunelli and D. Falavigna. Person identification using multiple cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10):955–966, October 1995.

[6] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, Oct 1993.

[7] J.P Campbell. Speaker recognition: A tutorial. *Proceedings of IEEE*, 85:1437–1462, 1997.

[8] R. Chellappa, C.L. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proceedings of the IEEE*, 83(5):705–740, May 1995.

[9] G. W. Cottrell and M. Fleming. Face recognition using unsupervised feature extraction. In *Proceedings Int. Neural Network Conference*, volume 1, pages 322–325, 1990.

[10] R.A Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(II):179–188, 1936.

[11] L. Hong and A.K Jain. Integrating faces and fingerprint for personal identification. *IEEE Trans. PAMI*, 20(12), 1997.

[12] A.K Jain, R. Bolle, and S. Pankanti. *Biometrics: Personal Identification in Networked Society.* Kluwer Academic Publishers, 1998.

[13] A.K Jain, L. Hong, and R. Bolle. On-line fingerprint verification. *IEEE Trans. PAMI*, 19(4), 1997.

[14] A.K Jain, L. Hong, and Y. Kulkarni. A multimodal biometric system using fingerprints, face and speech. In *2nd Intl. Conference on Audio-Video based Biometric Person Authentication*, 1999.

[15] M. Kirby and L. Sirovitch. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Trans. PAMI*, 12(1):103–108, 1990.

[16] J. Kittler, M. Hatef, R.P.W Duin, and J. Matas. On Combining Classifiers. *IEEE PAMI*, 20(3):226–239, 1998.

[17] J. Kittler and A Hojjatoleslami. A weighted combination of classifiers employing shared and distinct representations. In *Proc. Conference on CVPR*, pages 924–929, 1998.

[18] J. Kittler. Combining classifiers: A theoretical framework. *Pattern Analysis and Applications*, 1:18–27, 1998.

[19] M. Lades, J. Vorbruggen, J. Buhmann, J. Lange, C. V. D. Malburg, and R. Wurtz. Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. Comput*, 42:300–311, 1993.

[20] H.C Lee and R.E Gaensslen. *Advances in Fingerprint Technology.* Elsevier New-York, 1991.

[21] J. Lüttin and G. Maître. Evaluation protocol for the extended m2vts database (xm2vtsdb). Technical Report IDIAP-COM 98-05, IDIAP, 1998.

[22] J. Lüttin. Speaker verification experiments on the xm2vts database. Technical Report IDIAP-RR 99-02, IDIAP, 1999.

[23] J. L. Melsa and D. L. Cohn. *Decision and Estimation Theory.* McGraw-Hill, 1978.

[24] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. In *Early Visual Learning*, pages 99–130. Oxford University Press, 1995.

[25] XM2VTS multimodal database. Acts-m2vts project. Available at http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtdb.

[26] Penio S. Penev and Joseph J. Atick. Local Feature Analysis: A general statistical theory for object representation. *Network: Computation in Neural Systems*, 7(3):477–500, 1996.

[27] J.R Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, 1993.

[28] L.R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[29] A. E. Rosenberg, C. H. Lee, and S. Gokoen. Connected word talker verification using whole word hidden markov model. In *ICASSP-91*, pages 381–384, 1991.

[30] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[31] M. A. Turk and A. P. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3 (1):71–86, 1991.

[32] V. Vapnik. *Statistical Learning Theory*. Wiley Inter-Science, 1998.

[33] P .K. Varshney. *Distributed Detection and Data Fusion*. Springer, 1996.

[34] R. Viswanathan and P. K. Varshney. Distributed detection with multiple sensors. *Proceedings of the IEEE*, 85:54–63, January 1997.

[35] P. Werbos. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD thesis, Harvard University, 1974.

[36] J. Zhang, Y. Yan, and M. Lades. Face recognition: Eigenfaces, elastic matching, and neural nets. *Proceedings of IEEE*, 85:1422–1435, 1997.