# Bi-Modal Face and Speech Authentication: a BioLogin Demonstration System

Sébastien Marcel, Johnny Mariéthoz, Yann Rodriguez and Fabien Cardinaux
IDIAP Research Institute
Martigny, Switzerland 1920
marcel@idiap.ch
http://www.idiap.ch

## Abstract

*This paper presents a bi-modal (face and speech) authentication demonstration system that simulates the login of a user using its face and its voice. This demonstration is called BioLogin. It runs both on Linux and Windows and the Windows version is freely available for download. BioLogin is implemented using an open source machine learning library and its machine vision package.*

## 1. Introduction

Biometric identity authentication systems are based on the characteristics of a person, such as face, voice, fingerprint, iris, gait, hand geometry or signature. Identity authentication using the face or the voice information is a challenging research area that is currently very active, mainly because of the natural and non-intrusive interaction with the authentication system. An identity authentication system has to deal with two kinds of events: either the person claiming a given identity is the one who he claims to be (in which case, he is called a *client*), or he is not (in which case, he is called an *impostor*). Moreover, the system may generally take one decision: either *accept* the *client* or *reject* him and decide he is an *impostor*.

Biometrics have poor reputation because they are still not good enough for security and can be defeated [26]. The main drawback of biometrics is that if your biometric is stolen it is for life. However, it is possible to circumvent this problem if you can verify that the biometric came from the person at the time of the authentication and if you use multiple biometrics. Indeed, it is always possible to attack a biometric system using mimicry or pictures/recordings of some kind. Thus, there is a need to ensure that the biometric reading is contemporary and correlates multiple sources. It has been shown that the use of multiple modalities increases the performance of biometric systems. Most of these multimodal biometric systems perform fusion and sometime take advantage of temporal correlations between modalities. Indeed, very little work in the research community has been done on joint multi-modal fusion [3] to authenticate several modalities (for instance face and speech) at the same time.

In this paper, we present a bi-modal (face and speech) authentication demonstration system that simulates the login of a user using its face and its voice. This demonstration is called BioLogin. It runs both on Linux and Windows and the Windows version is freely available for download. BioLogin is implemented using an open source machine learning library and its machine vision package. Both the face and the speaker authentication system are based on the same statistical framework: Gaussian Mixture Models.

The paper is organized as follows. We first introduce the reader to the state-of-the-art in face and speaker authentication. Then we shortly present the approach implemented in the demonstrator. Next, we provide experimental results obtained by the algorithms on two well-known benchmark databases, namely XM2VTS and BANCA. Finally, we present the demonstration system and we conclude.

## 2. Face and Speaker Authentication

Most of face and speaker authentication systems are preceded by a segmentation procedure. This is a difficult task depending on the quality of the capture device, the conditions (illumination, complex background, noisy environment) and of the cooperation of the subject (face pose, occlusion, and clean speech). Segmentation is a general task in signal processing which consists of extracting relevant information (the face region in an image or speech frames in an audio signal) and filtering out irrelevant information (the background of an image, the silence or noise in the audio signal).

## 2.1. Face Authentication

### 2.1.1. Face Localization

The reliability and response time of face localization has a major influence on the performance and usability of subsequent processing such as face authentication. The goal of face detection/localization is to locate human faces in images at different positions, scales, orientations and lighting conditions. Face localization is a simplified face detection problem with the assumption that the image contains one and only one face.

In the past five years, face detection has been very popular in the computer vision research community, but it still remains a fundamental problem in pattern recognition. It is a difficult task because faces are non-rigid, dynamic objects with a high variability in shape, color and texture. Moreover face detection must be able to handle faces under various lighting conditions, orientation and pose.

A lot of methods have been proposed to solve frontal, and more recently non-frontal face detection. Among all these approaches, machine learning algorithms such as Support Vector Machines (SVMs) [32], Neural Networks [36], Bayesian classifiers [9], Hidden Markov Models [31] or boosting algorithms [33] have received much attention and shown outstanding results. In this project, we will only consider the most popular and efficient methods reported in the literature. For a more exhaustive survey, see the very complete paper of Yang et al. [39].

### 2.1.2. Face Authentication

Face recognition, authentication and identification are often confused. Face recognition is a general topic that includes both face identification and face authentication (also called verification). On one hand, face authentication is concerned with validating a claimed identity based on the image of a face, and either accepting or rejecting the identity claim (one-to-one matching). On the other hand, the goal of face identification is to identify a person based on the image of a face. This face image has to be compared with all the registered persons (one-to-many matching).

The problem of face authentication has been addressed by different researchers using various approaches. Thus, the performance of face authentication systems has steadily improved over the last few years. For a survey and comparison of different approaches see [8, 40, 28]. These approaches can be divided into *discriminant* approaches and *generative* approaches.

- *Discriminant Approaches*: A discriminant approach takes a binary decision (whether or not the input face is a client) and considers the whole input for this purpose. Such *holistic* approaches are using the original gray-scale face image or its projection onto a Principal Component subspace (referred to as PCA or Eigenfaces [38]) or Linear Discriminant subspace (referred to as LDA or Fisherfaces [1, 11]) as input of a discriminant classifier such as Multi-Layer Perceptrons (MLPs) [23, 22], Support Vector Machines (SVMs) [18] or simply a metric [20, 17].

- *Generative Approaches*: Recently, it has been shown that generative approaches such as Gaussian Mixture Models (GMMs) [7] and Hidden Markov Models (HMMs) [30, 29, 13, 6] were more robust to automatic face localization than the above discriminant methods. A generative approach computes the likelihood of an observation (a holistic representation of the face image) or a set of observations (local observations of particular facial features) given a client model and compares it to the corresponding likelihood given an impostor model.

Finally, the decision to accept or reject a claim depends on a score (distance measure, MLP output or Likelihood ratio) which could be either above (accept) or under (reject) a given threshold.

During recent international competitions on face authentication [27, 7], it has been shown that the discriminant approaches perform very well on manually localized faces. Unfortunately, these methods are not robust to automatic face localization (imprecision in translation, scale and rotation) and their performance degrades. On the opposite, generative approaches emerged as the most robust methods using automatic face localization.

## 2.2. Speech Authentication

### 2.2.1. Speech/Silence Detection

Speech/silence detection consists in isolating speech frames (relevant information for speaker authentication) from the rest of the audio signal. In any given speech sentence, silence often appears between words. These silence segments obviously do not contain much speaker information. Hence, state-of-the-art speaker authentication systems usually remove them with the help of a silence/speech detector. In fact, the main reason to remove them is that they influence the overall score: the more there are silence frames that are not removed, the smaller will be the amplitude of the score after normalization. Hence, since this score is then compared to a fixed threshold to take a decision, the underlying system becomes sensitive to the number of silence frames, which should be avoided.

#### 2.2.2. Speaker Authentication

The goal of a speaker authentication system is to decide whether a given *speech utterance* has been pronounced by a claimed client or by an impostor. A good introduction to the field can be found in [15, 4]. Different scenarios can take place in this framework, mainly *text dependent* and *text independent* speaker authentication, but they all use the same general statistical framework.

In this framework, one first needs a probabilistic model of *anybody*'s voice, often called a *world model* and trained on a large collection of voice recordings of several people. From this generic model, a more specific, client-dependent model, is then derived using adaptation techniques, using data from a particular client. One can then estimate the ratio of the likelihood of the data corresponding to some access with respect to the model of the claimed client identity, with the likelihood of the same data with respect to the *world model*, and accept or reject the access if the likelihood ratio is higher or lower than a given threshold, selected in order to optimize either a low rejection rate, a low acceptance rate, or some combination of both.

- In the context of *text independent* speaker authentication systems, where the trained client model would in theory be independent of the precise sentence pronounced by the client, the most used class of models is the Gaussian Mixture Model (GMM) with diagonal covariance matrix, adapted from a *world model* using MAP adaptation techniques [35].

- In *text dependent* speaker authentication, the system associates a sentence to each client speaker. Indeed, a possible solution to avoid replay attacks using speech recordings is to instruct the user to speak random words/digits at the time of the authentication. During an access, a client needs to say his associated sentence, which is known by the system. Therefore, the model created for each speaker can use the lexical information of the sentence in order to be more client and text specific. Models known to efficiently use this lexical information, such as Hidden Markov Models (HMMs) [34], need more resources (in space and time, during enrollment and test) than text independent models.

### 2.3. Bi-Modal Authentication

In the past 10 years, it has been shown that combining biometric authentication systems [19] achieves better performance than techniques using only one biometric modality (based on the face and the voice of an individual). This has been shown to be true using various *fusion* algorithms.

- *Fusion*: Fusion algorithms are methods whose goal is to merge the prediction of many algorithms (multiple biometric modules) in the hope of a better average performance than any of the individual methods. This fusion can be simple (maximum score, product or sum rules), but it is often better to train a fusion system using Machine Learning algorithms [5] such as MLPs or SVMs.

- *Joint Bi-Modal Authentication*: Recently, Asynchronous HMM [3] have been proposed for the task of bi-modal authentication [2]. This model specifically takes into account temporal correlations jointly between the audio and video streams, allowing for resynchronization between the streams. Often, lip movements do not appear at the same time the sound is uttered. It is possible to take this correlation into account by stretching or compressing streams with respect to each other.

## 3. The Proposed Approach

In this section, we describe IDIAP face and speaker authentication systems [7, 24]. Video and audio streams capture and processing are performed separately (Fig. 1). However, face and speaker authentication are based on the same statistical framework. The main difference lies in the specific image and audio features.

### 3.1. Face Segmentation

#### 3.1.1. Face Localization

The face localization system is based on the cascade paradigm of [33] and also on the use of Modified Census Transform (MCT) features [14]. MCT belongs to the family of Local Binary Pattern (LBP) features. By contrast to the Haar-like features used by Viola and Jones, LBP features are invariant to illumination and summarizes the local structure of the image.

Like most face detection systems, the face detector scans the input image at many scales. The conventional approach is to compute a pyramid of sub-sampled images like Rowley et al. [36]. A fixed scale sub-window then scanned across each of these images and sent to the cascade.

#### 3.1.2. Feature Extraction

The face image ($64 \times 80$ pixels) is decomposed in terms of $8 \times 8$ overlapping blocks. Then, Discrete Cosine Transform (DCT) is applied to every block and a sequence ($X_1^F = \{\mathbf{x}_1 ... \mathbf{x}_F\}$) of DCTmod2 frames is computed. DCTmod2 [37] frames are built from DCT frames (15 DCT coeff. $-3$ first coeff. $+3\ \Delta_x\ +3\ \Delta_y$). Thus, $\mathbf{x}_f \in \mathbb{R}^{18}$.

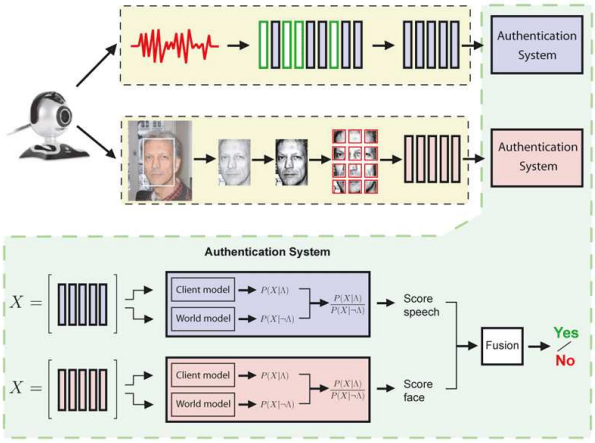


**Figure 1. IDIAP face and speaker authentication systems.**

## 3.2. Speech Segmentation

### 3.2.1. Speech/Silence Detection

The simplest approach to remove silences is to compare the energy level of each frame to a given threshold learned a priori on a separate training set. The threshold could also be learned or adapted on the first few frames which are hypothesized to be silence. Unfortunately, this is not always true; in real cases, it may happen that the first frames contain speech. Hence, a better strategy [25, 21] learns in an unsupervised way a bi-Gaussian model with the hypothesis that the distribution of the silence parts should be different from that of the speech part. The hypothesis that the log energy coefficient of the speech is bigger than the silence one is used to label each of the two Gaussians. Afterward, all the frames such that their probability under the *speech* Gaussian are smaller than their probability under the *silence* Gaussian are removed.

### 3.2.2. Feature Extraction

The speech signal is decomposed into a sequence of Linear Freq Cepstral Coefficient (LFCC) frames. These LFCC frames are expanded with their derivatives and log energy derivatives. This produces a sequence ($Y_1^S = \{\mathbf{y}_1 ... \mathbf{y}_S\}$) of frames, where $\mathbf{y}_s \in \mathbb{R}^{33}$.

## 3.3. Authentication

As stated above, face and speaker authentication are based on the same statistical framework.

Let us denote the parameter set for client $C$ as $\lambda_C$, and the parameter set describing a generic non-client as $\neg\lambda_C$. Given a claim for client $C$'s identity and a set of feature vectors $X$ supporting the claim, we find an opinion $\Lambda(X)$ on the claim using:

$$\Lambda(X) = \log P(X|\lambda_C) - \log P(X|\neg\lambda_C) \quad (1)$$

where $P(X|\lambda_C)$ is the likelihood of the claim coming from the true claimant and $P(X|\neg\lambda_C)$ is the likelihood of the claim coming from an impostor.

Finally, the decision to accept or reject a claim depends on the score $\Lambda(X)$ which could be either above (accept) or under (reject) a given threshold.

### 3.3.1. Enrollment

We can use different ways to train each client model. Traditional Maximum Likelihood training, such as Expectation-Maximization, can be used [10, 12]. Maximum A Posteriori (MAP) training [16] can also be used to adapt a generic model using client data. This MAP strategy was chosen because this approach is able to deal with a small amount of training data.

MAP training consists in:

1. training a world model $\neg\lambda_C$ from a large dataset by Maximum Likelihood,
2. adapting a client model $\lambda_C$ from $\neg\lambda_C$ using client data by Maximum A Posteriori.

### 3.3.2. Test

The above probabilities (Eq. 1) are represented by diagonal Gaussian Mixture Models. Each face model is a diagonal GMM ($\lambda^f$ and $\neg\lambda^f$) with $512$ gaussians ($18'944$ parameters). And each speech model is a diagonal GMM ($\lambda^s$ and $\neg\lambda^s$) with $200$ gaussians ($13'400$ parameters).

Then, the respective face and speech scores are computed using Eq. 2 and Eq. 3.

$$\Lambda_C^f(X_1^F) = \log P(X_1^F|\lambda_C^f) - \log P(X_1^F|\neg\lambda_C^f) \quad (2)$$
$$\Lambda_C^s(Y_1^S) = \log P(Y_1^S|\lambda_C^s) - \log P(X_1^S|\neg\lambda_C^s) \quad (3)$$

### 3.4. Fusion

The goal of fusion is to merge outputs of face and speech experts (2 or more) into a feature vector $[\Lambda^1(X), ..., \Lambda^n(X)]$ and try to classify it as a client or an impostor. This can be achieved using a classifier. In our framework, we decided to use a simple linear classifier:

$$P(X, Y|C) = w \cdot \Lambda_C^f(X) + (1 - w) \cdot \Lambda_C^s(Y) \quad (4)$$

Finally, fusion produces an opinion $\Lambda^*(X.Y)$ that might be used for final decision.

## 4. Experiment results

The machine learning library used for all experiments is **Torch** and its machine vision package **Torch vision**. More details are provided in Section 5.

### 4.1. Databases

We performed face and speaker authentication experiments on two well-known multimodal databases, namely XM2VTS and BANCA.

#### 4.1.1. XM2VTS

The XM2VTS database [1] contains synchronized image and speech data recorded on 295 subjects during four sessions taken at one month intervals. The 295 subjects were divided, according to the *Lausanne Protocol*, into a set of 200 clients, 25 evaluation impostors, and 70 test impostors. Two different evaluation configurations were defined. They differ in the distribution of client training and client evaluation data. We performed the experiments following the *Lausanne Protocol Configuration I*.

#### 4.1.2. BANCA

The BANCA database [2] was designed in order to test multimodal identity authentication with various acquisition devices (2 cameras and 2 microphones) and under several scenarios (controlled, degraded and adverse). For 5 different languages (English, French, German, Italian and Spanish), video and speech data were collected for 52 subjects (26 males and 26 females), i.e. a total of 260 subjects. Each language - and gender - specific population was itself subdivided into 2 groups of 13 subjects (denoted $g1$ and $g2$). Each subject participated to 12 recording sessions, each of these sessions containing 2 records: 1 true *client access* (T) and 1 informed [3] *impostor attack* (I). For the image part of the database, there is 5 shots per record. The 12 sessions were separated into 3 different scenarios.

In the BANCA protocol, we consider that the true client records for the first session of each condition is reserved as training material. In all our experiments, the client model training is done on at most these 3 records. We consider the following protocols, namely Matched Controlled (Mc) and Pooled test (P) protocol, where one controlled session is used for client training and, the same controlled conditions sessions for Mc, and all conditions sessions for P, are used for client and impostor testing.

### 4.2. Performance Evaluation

The authentication decision is then reached as follows. Given a threshold $\tau$, the claim is accepted when $\Lambda^*(X, Y) \geq \tau$, and is rejected when $\Lambda^*(X, Y) < \tau$. This threshold is chosen to optimize a given criterion such as the Equal Error Rate ($EER$), i.e when $FAR = FRR$ (Fig. 2).

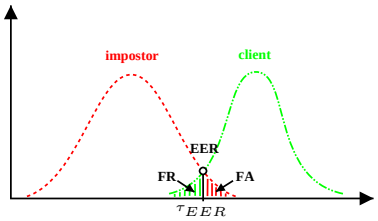


**Figure 2. Illustration of typical errors of a biometric system.**

$FRR$ is the False Rejection Rate (when the system rejects a client), $FAR$ is the False Acceptance Rate (when the system accepts an impostor), $HTER$ is the Half Total Error Rate (an unique measure given by $HTER = \frac{FRR+FAR}{2}$).

### 4.3. Results

We present baseline results (in terms of HTER), on XM2VTS and BANCA databases, obtained by IDIAP systems (Table 1). In order to provide an unbiased evaluation of the performance, the decision threshold has to be chosen

[1] http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb
[2] http://www.ee.surrey.ac.uk/banca
[3] The actual speaker knew the text that the claimed identity speaker was supposed to utter.

a priori (not optimize on the test set itself). Thus, we determine the threshold $\tau$ on the development set which minimizes the $EER$ criterion.

| | XM2VTS (LP1) | BANCA (Mc) | BANCA (P) |
|---|---|---|---|
| Face | 1.67 | 5.77 | 18.96 |
| Speech | 1.14 | 4.32 | 12.29 |
| Fusion | 0.48 | 4.32 | 9.99 |

**Table 1. Baseline results obtained by IDIAP systems in terms of $HTER$.**

For experiments on XM2VTS database, we use all available training client images to build the generic face model and additional set of data to build the generic speech model. For BANCA experiments, the generic model was trained with the additional set of data, referred to as *world data* (independent of the subjects in the client database).

## 5. BioLogin Demonstration System

Several demonstration systems of state-of-the-art technologies in person authentication can be found on the internet. However, to our knowledge there exist no demonstration systems of bi-modal face and speaker authentication freely available and implemented using open source libraries.

### 5.1. Description

The BioLogin bi-modal authentication system is freely available [4]. BioLogin requires Windows XP (SP 1 or SP 2), DirectX 9.0b and a Logitech camera. It has been tested using QuickCam Pro 4000, QuickCam for Notebooks Pro, QuickCam Zoom, QuickCam Orbit/Sphere.

The system (Fig. 3) includes two applications:

- *BioLogin*: login using the face and the voice (test a biometric template),
- *User Manager*: creates a new account and enables the user to enroll a bi-modal biometric template.

First the user needs to create his/her account using the Manager application. The registration consists in (1) filling a form and (2) recording a session of four audio/video shots. During each shot, the system asks the user to pronounce his/her pass-phrase. The audio recording starts when a face is detected and stops when the time is elapsed or when the user press `<enter>`. Face images are automatically captured during the audio recording. At the end of the recording session, the user can visualize/listen to the recordings. The user can decide to cancel the recording session and to perform another one or to enroll his/her model from recordered data. The enrollment process takes only few seconds.

Finally, the user can launch the BioLogin application. This application presents a list of registered persons. To perform an authentication test, the user simply needs to select a person. Then the audio/video capture is immediately launched. As soon as the face is detected, the user has a few seconds to pronounce the pass-phrase. If the time is elapsed or if the user press `<enter>` then the authentication is performed. The system displays either **accepted** in green if the user is considered as a client or **rejected** in red if the user is considered as an impostor. The authentication process is very fast and it is therefore possible to perform many true-client accesses or impostor accesses by choosing a different registered person.

### 5.2. Open Source Software

The BioLogin is based on two open source libraries:

- **Torch** [5] is a machine-learning library developed at IDIAP. It is written in simple C++ and distributed under a BSD license. Torch implements a lot of things in gradient machines (multi-layered perceptrons, radial basis functions, mixtures of experts, convolutional networks, ...), Support vector machines (in classification and regression), Ensemble models such as bagging or adaboost, Non-parametric models such as K-nearest-neighbors, Distributions such as Kmeans, Gaussian Mixture Models, Hidden Markov Models, Input-Output Hidden Markov Models, and Speech recognition tools (Embedded training and large vocabulary decoding).

- **Torch vision** [6] is a machine vision library also developed at IDIAP and based on **Torch**. Torch vision provides basic image processing and feature extraction algorithms such as rotation, flip, photometric normalizations (Histogram Equalization, Multi-scale Retinex, Self-Quotient Image or Gross-Brajovic), edge detection, 2D DCT, 2D FFT, 2D Gabor, PCA, LDA. It provides also various metrics (Euclidean, Mahanalobis, ChiSquare, NormalizeCorrelation, TangentDistance, ...) and modules for face detection (MLP, cascades of Haar-like classifiers) and face recognition/authentication.

[4] http://www.idiap.ch/∼marcel/en/biometry.php

[5] http://www.torch.ch

[6] http://www.idiap.ch/∼marcel/en/torch3/introduction.php

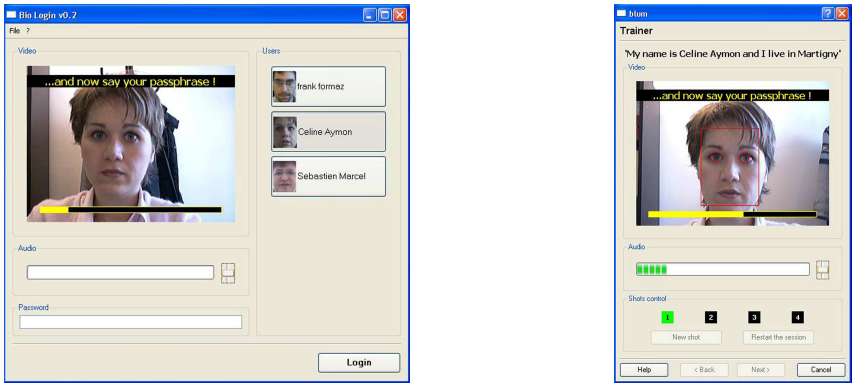


**Figure 3. Bi-Modal Authentication system based on face, speech and fusion developed at IDIAP. The system provides a BioLogin application (left) to test a client, and a Manager application (right) to create a new account by enrollment.**

## 6. Conclusion

In this paper, we presented a bi-modal (face and speech) authentication demonstration system that simulates the login of a user using its face and its voice. This demonstration is called BioLogin. It is based on Gaussian Mixture Models used both for face and speaker authentication. BioLogin is implemented using an open source machine learning library and its machine vision package. It runs both on Linux and Windows and the Windows version is freely available for download.

## Acknowledgments

The authors wish to thank the Swiss National Science Foundation for supporting this work through the National Center of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)". The authors would like also to thank Dr. Samy Bengio and Dr. Conrad Sanderson for useful comments and contributions.

## References

[1] P. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specifi c linear projection. In *ECCV'96*, pages 45–58, 1996. Cambridge, United Kingdom.

[2] S. Bengio. Multimodal authentication using asynchronous HMMs. In *4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA*, pages 770–777. Springer-Verlag, 2003.

[3] S. Bengio. Multimodal speech processing using asynchronous hidden markov models. *Information Fusion*, 5(2):81–89, 2004.

[4] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovsk-Delacrtaz, and D. Reynolds. A tutorial on text-independent speaker verifi cation. *EURASIP Journal on Applied Signal Processing*, 4:430–451, 2004.

[5] C. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

[6] F. Cardinaux, C. Sanderson, and S. Bengio. Face verifi cation using adapted generative models. In *The 6th International Conference on Automatic Face and Gesture Recognition*, Seoul, Korea, 2004. IEEE.

[7] F. Cardinaux, C. Sanderson, and S. Marcel. Comparison of MLP and GMM classifi ers for face verifi cation on XM2VTS. In *Proceedings of the International Conference on Audio- and Video-Based Biometric Person Authentication*. Springer-Verlag, 2003.

[8] R. Chellappa, C. Wilson, and C. Barnes. Human and machine recognition of faces: A survey. Technical Report CAR-TR-731, University of Maryland, 1994.

[9] T. Cootes, K. Walker, and C. Taylor. View-based active appearance models. In *Proceedings of the IEEE Conference on Automatic Face and Gesture Recognition*, pages 227–232, 2000.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society B*, 39:1–38, 1977.

[11] P. A. Devijver and J. Kittler. *Pattern Recognition: A Statistical Approach*. Prentice-Hall, Englewood Cliffs, N.J., 1982.

[12] R. Duda, P. Hart, and G. Stork. *Pattern Classification*. 2001.

[13] S. Eickeler, S. Müller, and G. Rigoll. High Performance Face Recognition Using Pseudo 2D-Hidden Markov Models. In *European Control Conference (ECC)*, Karlsruhe, Germany, 1999.

[14] B. Froba and A. Ernst. Face detection with the modifi ed census transform. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition (AFGR)*, Seoul, Korea, May 2004.

[15] S. Furui. Recent advances in speaker recognition. In Springer, editor, *Audio- and Video-based Biometric Person Authentication*, pages 237–252, 1997.

[16] J.-L. Gauvain and C.-H. Lee. Maximum *a posteriori* estimation for multivariate gaussian mixture observations of markov chains. *IEEE Trans. Speech and Audio Processing*, 2(2):291–298, 1994.

[17] J. K. J, R. Ghaderi, T. Windeatt, and G. Matas. Face verification via ECOC. In *British Machine Vision Conference (BMVC01)*, pages 593–602, 2001.

[18] K. Jonsson, J. Matas, J. Kittler, and Y. Li. Learning support vectors for face verification and recognition. In *4th International Conference on Automatic Face and Gesture Recognition*, pages 208–213, 2000.

[19] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[20] Y. Li, J. Kittler, and J. Matas. On matching scores of LDA-based face verification. In T. Pridmore and D. Elliman, editors, *Proceedings of the British Machine Vision Conference BMVC2000*. British Machine Vision Association, 2000.

[21] I. Magrin-Chagnolleau, G. Gravier, and R. Blouet. Overview of the 2000-2001 ELISA consortium research activities. In *A Speaker Odyssey*, pages 67–72, Chania, Crete, Greece, June 2001.

[22] S. Marcel. A symmetric transformation for lda-based face verification. In *Proceedings of the 6th International Conference on Automatic Face and Gesture Recognition*. IEEE Computer Society Press, 2004.

[23] S. Marcel and S. Bengio. Improving face verification using skin color information. In *Proceedings of the 16th ICPR*. IEEE Computer Society Press, 2002.

[24] J. Mari´ethoz and S. Bengio. A comparative study of adaptation methods for speaker verification. In *International Conference on Spoken Language Processing ICSLP*, pages 581–584, Denver, CO, USA, September 2002.

[25] J. Mari´ethoz and S. Bengio. An alternative to silence removal for text-independent speaker verification. Technical Report IDIAP-RR 03-51, IDIAP, 2003.

[26] T. Matsumoto, H. Matsumoto, K. Yamada, and S. Hoshino. Impact of artificial gummy fingers on fingerprint systems. *Proceedings of SPIE (Optical Security and Counterfeit Deterrence Techniques IV)*, 4677, 2002.

[27] K. Messer, J. Kittler, M. Sadeghi, M. Hamouz, A. Kostyn, S. Marcel, S. Bengio, F. Cardinaux, C. Sanderson, N. Poh, Y. Rodriguez, J. Czyz, and al. Face authentication test on the BANCA database. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, Cambridge, August 23-26 2004.

[28] K. Messer, J. Kittler, M. Sadeghi, M. Hamouz, A. Kostyn, S. Marcel, S. Bengio, F. Cardinaux, C. Sanderson, N. Poh, Y. Rodriguez, K. Kryszczuk, J. Czyz, L. Vandendorpe, J. Ng, H. Cheung, and B. Tang. Face authentication competition on the BANCA database. In *Proceedings of the International Conference on Biometric Authentication (ICBA)*, Hong Kong, July 15-17 2004.

[29] A. Nefian and M. Hayes. Hidden markov models for face recognition. In *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 2721–2724, 1998.

[30] A. Nefian and M. Hayes. Face recognition using an embedded HMM. In *Proceedings of the IEEE Conference on Audio and Video-based Biometric Person Authentication (AVBPA)*, pages 19–24, 1999.

[31] A. Nefian and M. H. III. Face detection and recognition using hmm. In *Proceedings of the IEEE International Conference on Image Processing*, volume 1, pages 141–145, 1998.

[32] E. Osuna, R. Freund, and F. Girosi. Training svm: An application to face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 130–136, 1997.

[33] P. Viola and M. Jones. Robust Real-time Object Detection. In *IEEE ICCV Workshop on Statistical and Computational Theories of Vision*, 2001.

[34] L. Rabiner and B.-H. Juang. *Fundamentals of speech recognition*. Prentice All, first edition, 1993.

[35] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1–3), 2000.

[36] H. Rowley, S. Baluja, and T. Kanade. Rotation invariant neural network-based face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 38–44, 1998.

[37] C. Sanderson and K. K. Paliwal. Fast features for face authentication under illumination direction changes. *Pattern Recognition Letters*, 14(24), 2003.

[38] M. Turk and A. Pentland. Eigenface for recognition. *Journal of Cognitive Neuro-science*, 3(1):70–86, 1991.

[39] M.-H. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and machine Intelligence*, 24(1):34–58, January 2002.

[40] J. Zhang, Y. Yan, and M. Lades. Face recognition: Eigenfaces, Elastic Matching, and Neural Nets. In *Proceedings of IEEE*, volume 85, pages 1422–1435, 1997.