Addressing the Vulnerabilities of Likelihood-Ratio-Based Face Verification

Krzysztof Kryszczuk and Andrzej Drygajlo

Signal Processing Institute, Swiss Federal Institute of Technology Lausanne (EPFL)
{krzysztof.kryszczuk,andrzej.drygajlo}@epfl.ch

Abstract. Anti-spoofing protection of biometric systems is always a serious issue in real-life applications of an automatic personal verification system. Despite the fact that face image is the most common way of identifying persons and one of the most popular modalities in automatic biometric authentication, little attention has been given to the spoof resistance of face verification algorithms. In this paper, we discuss how a system based on DCT features with a likelihood-ratio-based classifier can be easily spoofed by adding white Gaussian noise to the test image. We propose a strategy to address this problem by measuring the quality of the test image and of the extracted features before making a verification decision.

1 Introduction

The goal of all automatic biometric verification systems is to reliably establish if the identity claim comes from the real claimant or from an impostor. Attempts to impersonate a selected person in order to gain privileges otherwise reserved for the rightful claimant, otherwise known as spoofing, have been not an unusual threat since personal identity verification became a necessity. Only quite recently systems that automatically compare voices, fingerprints, faces, irises and signatures, left the laboratories and met the challenges of the real world. One of those challenges is, and probably will remain, spoofing.

Moreover, more and more frequent are the attempts to store personal biometric information in a digital form and to embed this information in identity documents – like identity cards, passports, visas, company access cards, etc. One of the common biometric modality choices for those applications is face image. Digitally stored face images or templates are likely to soon accompany a traditional photograph, to allow both human and automated verification procedures.

The objective of the work presented in this paper is to show that an estimation of the quality of the test image is necessary to assure the robustness of a face verification system to spoofing. As defined in [8,10], a complete face verification system consists of modules that perform: 1) *localization*, 2) *normalization*, 3) *feature extraction*, 4) *classification*. We postulate to add an additional step before classification: *quality assessment*.

We show that omitting the quality assessment step may actually compromise the impermeability of an automated face verification system to imposters. Using an example of the local Discrete Cosine Transform (DCT)-feature based system with likelihood-ratio-based classifier, we show how an acceptable set of features can originate

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 426–435, 2005.
© Springer-Verlag Berlin Heidelberg 2005

from an alien signal (noise), which can successfully spoof a face verification system. Consequently, we propose to put additional constraints on the input signal in order to prevent such non-eligible accesses.

The paper is organized as follows: Section 2 gives an overview of features used for face verification. Section 3 focuses on face verification based on DCTmod2 features and Gaussian Mixture Model (GMM) classifier. Section 5 deals on how a discussed face verification system can produce unreliable verification decisions. Section 6 proposes two complementary quality assessment methods and their combination. Conclusions and future work prospects are found in Sections 7 and 8.

2 Feature Extraction Routines for Face Images

The most popular features for face recognition from 2D images can be divided into holistic and local [8,11,12]. The holistic features are probably in widest use. However, their recognition accuracy suffers from scaling, rotation and translation of the input signal [8,9].

Another group of feature extraction techniques is made of algorithms that divide the input image into segments and extract features from those segments independently, producing a set of feature vectors. Subject literature reports the use of local PCA [10], Gabor wavelets [8] and 2-dimensional DCT-based features and their derivates [4,5,8,9,10]. Local features reportedly suffer less than their holistic counterparts from incorrect geometrical normalization of the input signal, which manifests itself in good performance of modified DCT-based features, particularly in the tests involving automatically localized faces [6]. For this reason, we have chosen the local feature extraction approach, and a GMM-based classifier, as a testbed for our experiments.

To overcome the disadvantages of using only local or global feature extraction schemes alone, successful attempts have been made to create hybrid systems that use both approaches [4]. Although the overall performance of those systems is reported to be superior in comparison with non-hybrid approaches, they are also bound to suffer from attacks which would confuse one feature extraction scheme, leaving only the second one in operation.

3 DCTmod2-GMM Face Verification

In our experiments we used a face verification scheme implemented in similar fashion as presented in [5,8,10]. Images from BANCA database [2] (French part) were used to build the world model (520 images, 52 individuals, 338 Gaussians in the mixture), while images from BANCA (English part) database were used to build client models using a recursive adaptation of the world model, as described in [7]. The adaptation relevance parameter was set to 16, and the number of iterations was set to 10. The images used in the experiments were cropped, normalized and rescaled to the size of 64×64 pixels. All faces were localized manually and normalized geometrically (eye position). Mean pixel intensity subtraction was used as the data normalization procedure before feature extraction. More sophisticated normalization schemes grant slightly better verification performance [4], but minute gains in performance was not the objective of the experiments reported here.

To verify a claim that a given test image belongs to the client C, a set of feature vectors, X, is extracted from the image. The verification decision is based on the likelihood ratio:

$$LR(X) = \frac{L(X \mid \lambda_C)}{L(X \mid \lambda_W)} \tag{1}$$

where $L(X|\lambda_C)$ and $L(X|\lambda_W)$ are the joint likelihoods of the set of vectors X given λ_C (the model of client C), and λ_W (the *world model*) [8]. The value of $LR(X)$ is compared to a threshold Θ, whose value is computed depending on the desired properties of the verification system [2].

Making a decision based on the likelihood ratio was proved to be an optimal strategy for biometric verification based on fixed-length feature vectors [1]. This holds assuming that both λ_C and λ_W were built using representative data sample from their respective populations. In the case of face verification it means that λ_C and λ_W have to account for every possible condition and degree of quality of the input face image. In the case of the performance estimation based on a standard evaluation protocol (e.g. XM2VTS, BANCA) this condition is met. It may not be the case in a real-world application where there is no closed set of images that can appear as an identity claim.

If a significant mismatch exists between the quality of the test image and the quality of the images employed in the training of λ_C and λ_W, using the likelihood ratio stops to be a meaningful way of making reliable verification decisions. In a classical verification scheme, the only possible outcomes of the decision process are acceptance or rejection of the hypothesis that the claimant is who he claims to be. An image, whose quality does not match at all the quality of images used to train the models, cannot be correctly represented by those models. Therefore one could expect that upon encountering such an image, the system will reject the claim. In the likelihood ratio scheme though, if the world model explains the incoming data from the claimant to an even smaller extent, the decision of the system will be positive, which is an obviously meaningless result. We show that such situation is possible and quite likely in a real application.

4 Tricking a DCTmod2-GMM System

DCT-based local features capture predominantly higher spatial frequency in the image [4]. Therefore, in order to depart from image quality comparable (by means of the DCTmod2 features) to the quality of images used during the training of λ_C and λ_W, we corrupt the test images with white Gaussian noise. Such noise contamination introduces alien spatial frequencies to the image, and since the mean image intensity remains unchanged, the energy distribution between frequencies in the image alters.

We choose noise contamination as the way to depart from the initial image quality conditions because it is a likely factor to corrupt images in real life. The corruption was followed by normalization identical to that performed on the images used for the training of world and client models. Example images with different level of added noise can be found in Figure 1. Percentages of noise contamination of images are equivalent to noise-to-signal ratio (reciprocal of SNR).

Corrupted versions of face images have been prepared for all images from Sets 02, 03 and 04 of the English part of the BANCA database (total of 1560 images). Following tests were performed:

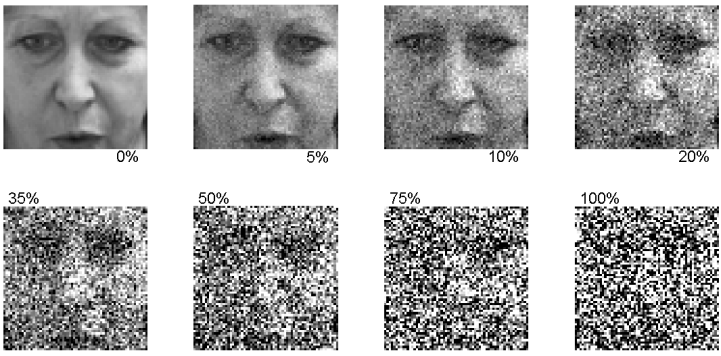

Fig. 1. Example face image from BANCA database (English part), corrupted with additive white Gaussian noise

- *Genuine client access tests*. Images corrupted with various amount of noise were tested using corresponding client models (corrupted images of client 1 against the model of client 1, etc.).
- *Impostor attack tests*. Images corrupted with various amount of noise were tested using client models created for another client. For simplicity, images coming from client n were tested against the model of client $n+1$. Face images of client 52 were attempting to impersonate client 1.

Genuine and impostor access attempts were scored using likelihood ratio $LR(X)$, as discussed in Section 3. The scores in terms of $LR(X)$ are presented in Figure 2. Gaussian approximations of their distributions are shown in Figure 3. For all tested images, $L(X|\lambda_C)$ was plotted against $L(X|\lambda_W)$ in Figure 4.

The influence of noise contamination on the likelihood scores is evident in Figure 4. For every X, the addition of noise causes a significant decrease of both $L(X|\lambda_C)$ and $L(X|\lambda_W)$, suggesting that the feature set X originating from the input image cannot be represented by neither the client, nor the world model. This information, however, is lost when $LR(X)$ is calculated (Figures 2 and 3). In this situation the verification system is bound to be confused.

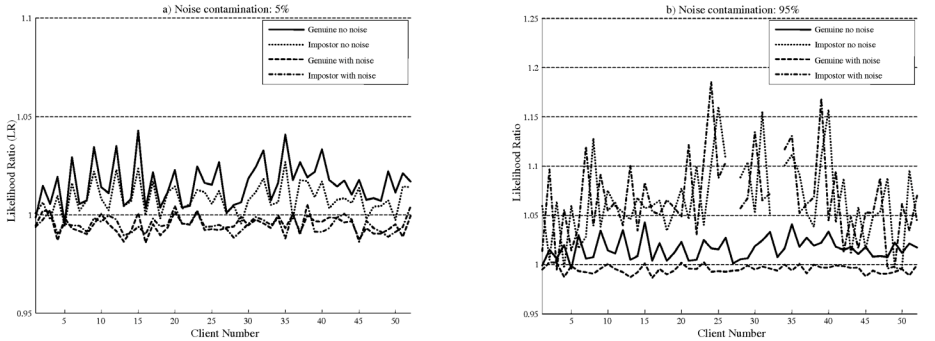

Fig. 2. Likelihood ratio scores for the verification tests on images from BANCA, English part, Session 03, for noise contamination 0% (no noise), and 5% and 95%

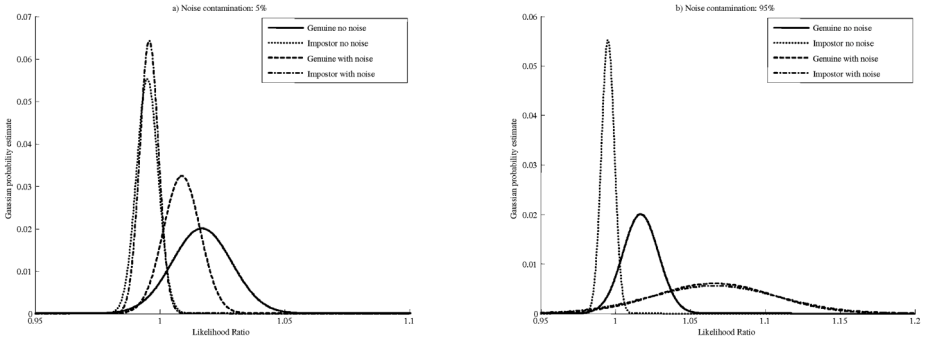

Fig. 3. Distributions of likelihood ratio scores for 0% (no noise), 5% and 95% of noise contamination

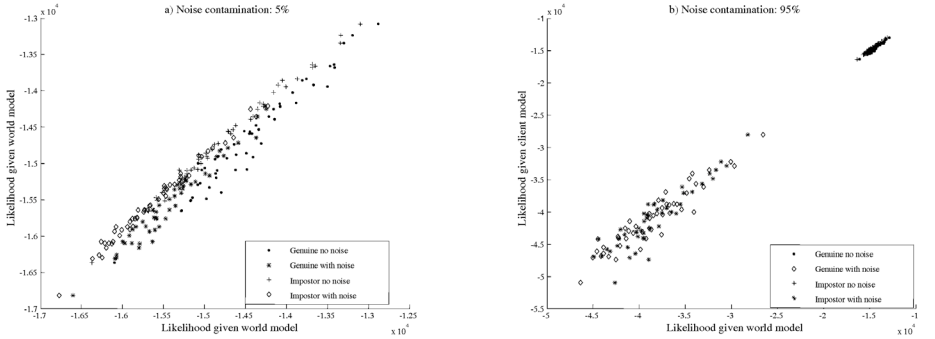

Fig. 4. Likelihood scores $L(X|\lambda_C)$ plotted against likelihood scores $L(X|\lambda_W)$

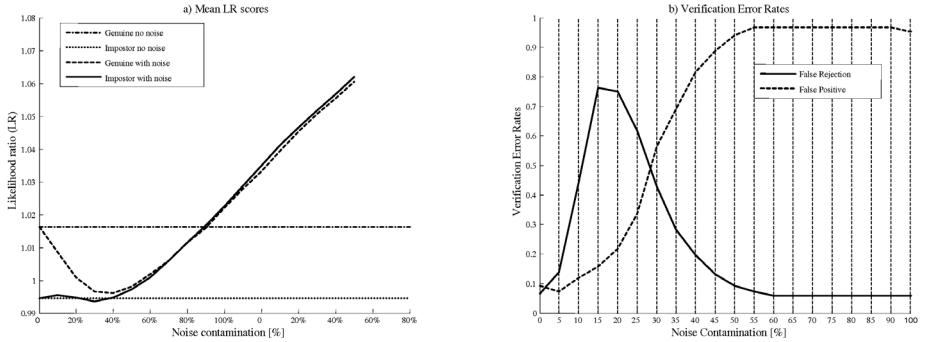

Fig. 5. a) Means of the score distributions of noise-contaminated genuine client and impostor claims. Scores for noise-free images left as a frame of reference, b) Mean verification results (BANCA, English, Sessions 02,03,04), as a function of noise contamination

The plots in Figure 5a represent the change of the mean distance between the genuine client and impostor likelihood ratio distributions, as a function of the noise contamination of face images.

As the presented results reveal, the automatic face verification system tends to reject impostors when the noise contamination is not significant. At those levels of

noise present in the input face images the average scores for real clients sink as well – a reasonable and desirable behavior which could be expected. For the same noise percentage, the impostor scores remain relatively constant. As the noise contamination of the input face image becomes large (above 25%) both genuine client and impostor scores grow rapidly, level up with the mean score for noise-free genuine clients at about 50% noise content, and continue growing. Figure 5b shows the verification error rates as a function of noise contamination (threshold Θ=1). Above 30% of noise contamination the system begins to favor acceptances over rejections, and above 50% of noise almost every claim is accepted.

5 Image Quality Assessment

Accepting every claim above certain level of image quality degradation is definitely an unacceptable behavior. Upon inspection of Figures 1 and 5b, it appears that the confused behavior of the system begins when the noise contamination begins to occlude the important facial features and the image bears less and less resemblance to a face. In order to address this vulnerability, it is necessary to introduce an intermediate step, which will automatically assess the quality of the input image. The goal of such assessment is:

1. To tell if the image presented to the system is indeed an image of a face.
2. To give a measure of the quality of the input image, relative to the quality of images used in the training of the system.

In order to meet those requirements, we consider two alternative approaches:

1. Quality assessment in the likelihood score domain.
2. Quality assessment independent of the features considered for verification.

5.1 Quality Assessment in the Likelihood Score Domain

The concept of likelihood-based verification, as expressed by Equation (1), is to find out if the feature vector is better represented by λ_C or by λ_W. This measure does not account for a situation when neither of the models represents the data adequately. We propose to compute a measure of how much the quality of the input matches either of the two models, or both simultaneously. For given feature set X originating from an image I we define the quality measure Q:

$$Q(I) = L(X \mid \lambda_C) + L(X \mid \lambda_W)\,. \tag{2}$$

The distribution of Q for N images I_T used in training of models λ_C and λ_W can be approximated using a mixture of 3 Gaussians, following identical model training procedure as during the training of λ_W. The distribution and the resulting model λ_Q are shown in Figure 6a. For given test image I we calculate its relative quality measure as:

$$R(I) = N\frac{L(Q(I) \mid \lambda_Q))}{\sum_{i=1}^{N} L(Q(I_T^i) \mid \lambda_Q)} \tag{3}$$

For every level of noise contamination of the n test images we calculate their corresponding mean relative quality measure $R_{mean}=(1/n)\cdot\sum R(I)$. Figure 6b shows R_{mean} as the function of the level of noise contamination.

The curve presented in Figure 6b is descending quickly from high relative quality values for clean and little noise-contaminated images, to arrive at values near zero for test images contaminated with more than 10% of white Gaussian noise.

The estimate is hence very sensitive to the degradation of the input image quality. At the same time, however, it depends heavily on the training conditions of λ_C and λ_W. Also, it really says nothing if the input image I is indeed a face image.

5.2 Quality Assessment Independent of the DCTmod2 Features

Upon inspection of Figure 1 one can conclude that gradual degradation makes first the individual facial features difficult to recognize, then even the rudimentary features stop to be obvious, until the image ceases to resemble a face at all. Since image quality should not be individual-dependent, it is desirable to have a measure of "face-likeness", in other words to estimate how much the input image resembles a face at all.

For this purpose, we propose to use normalized correlation of the input image with an average face template. We build the average face template T_F out of the same image set that was used before to build the world model λ_W, as described in [3]. The template can be seen in Figure 7a.

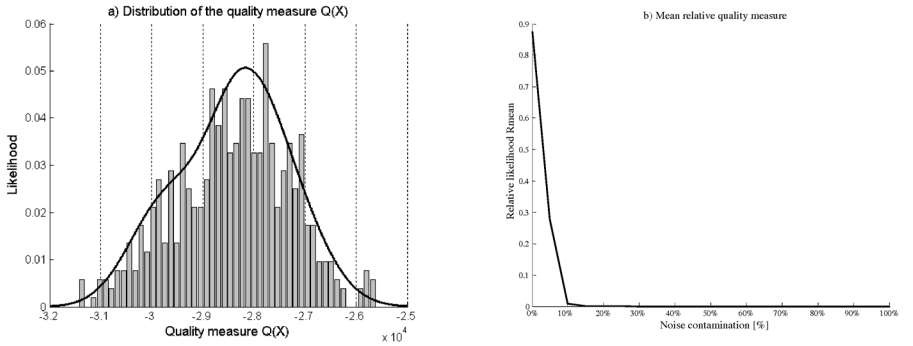

Fig. 6. a) Distribution of *Q(I)* and its corresponding GMM, b) Relative mean quality measure R_{mean} as a function of noise contamination

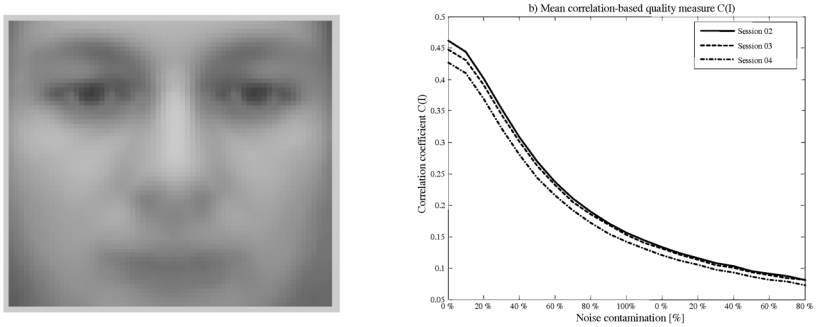

Fig. 7. Average face template T_F and b) mean correlation-based quality scores for images from BANCA (English), Sessions 02, 03 and 04

Since the average face template is a smooth reconstruction from first 8 principal components, it preserves only the facial features that are common to all faces from the training set. This makes it a good frame of reference to assess the "face-likeness" of an image. For given test image I we define its degree of resemblance to a face as:

$$C(I) = \max(corr(I, T_F)), \tag{4}$$

where $corr(I,T_F)$ is a normalized 2D correlation of T_F and I. Figure 7b shows how $C(I)$ changes as the function of noise contamination of the face image. The correlation-based quality measure gives a very good estimate if the input image indeed is a face image, independently of the features extracted for verification purposes.

Proposed correlation-based measure of "face-likeness" is one of the methods used in face detection [3]. Face detection, in general, is a way of assessing how much given object resembles a face. Therefore, in theory any face detection algorithm at some point does calculate some measure of "face-likeness" and this information can be used during the quality assessment step.

5.3 Combining Quality Measures for Increased Robustness

The relative quality measures $R(I)$ and $C(I)$ have complementary strengths and weaknesses. While $R(I)$ is more sensitive to the degradation of I in terms of features used for verification, $C(I)$ is providing information about how likely it is that I is an image of a face. Since both measures are computed independently, and it is required that an image of the rightful claimant is both an image of a face and that its quality is compatible with λ_C and λ_W, we define the combined quality measure $M(I)$ *a*s:

$$M(I) = R(I) \cdot C(I) \tag{5}$$

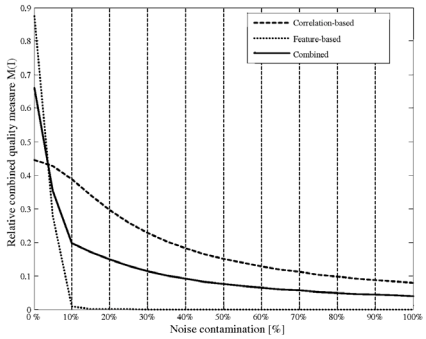

Fig. 8. Correlation-based, feature-based and combined relative quality measure of input face image, as a function of the percentage of noise contamination

Figure 8 presents the $M(I)$ as the function of the percentage of noise contamination.

Let's introduce a threshold Θ_R. For given test image I, if $M(I) < \Theta_R$, the quality assessment module rejects the image on the basis of its insufficient quality relative to the images used for the training of the verification system. The choice of Θ_R depends on the desired properties of the system. For example, if an increase of false acceptances is not desired, by comparison of curves in Figures 5b and 8, a threshold Θ_R=0.2 would be appropriate.

6 Summary and Conclusions

In this paper we have shown that a DCT-based face verification systems that uses likelihood-ratio-based classifier, can be vulnerable to spoofing attacks using face images contaminated with white Gaussian noise. We presented a method of obtaining an automatic assessment of the quality of face images which can help in preventing such attacks. The quality assessment method uses a combined measure that takes into account both the compatibility of the input image with world and client models, and the "face-likeness" of the image. The latter is particularly necessary in systems based on local features modeled with GMMs, since the spatial relations between facial features are not preserved in the models.

7 Future Work

We presented a combined quality assessment scheme for face images. The hybrid approach of this method gives a good global estimate of the quality of the image on the input of a face verification system. This estimate, however, gives no information as to why the quality is deteriorated, relative to the reference images. We are currently developing a set of techniques that allow a precise estimation of various quality measures of face images (localization, lighting, sharpness, etc.).

Acknowledgements

This work was supported in part by the Swiss National Science Foundation (SNSF) through the National Network of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM)2.

References

1. A.M. Bazen and R.N.J.Veldhuis.:Likelihood-Ratio-Based Biometric Verification. In: IEEE Transactions on Circuits and Systems for Video Technology, Vol. 14, No.1, January 2004.
2. S. Bengio, F. Bimbot, J. Mariethoz, V. Popovici, F. Por´ee, E. Bailly-Balliere, G. Matas and B. Ruiz.: Experimental protocol on the BANCA database. Technical Report IDIAP-RR 02-05, IDIAP, 2002. (www.idiap.ch)
3. K. Kryszczuk and A. Drygajlo.: Color Correction For Face Detection Based on Human Visual Perception Metaphor. In: Proc. of the Workshop on Multimodal User Authentication, p. 138-143, Santa Barbara, CA USA, 2003.
4. S. Lucey.: The Symbiotic Relationship of Parts and Monolythic Face Representations in Verification. In: International Workshop on Face Processing in Video (FPIV), Washington D.C., 2004.
5. S. Lucey and T. Chen.: A GMM Parts Based Face Representation for Improved Verification through Relevance Adaptation. In: Proc. of the IEEE CSS Conf. on Computer Vision and Pattern Recognition, Vol. 2, pp. 855-861, Washington, USA, 2004.
6. 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: Proc. of the International Conference on Biometric Authentication, ICBA, Hong Kong, 2004.

7. D.A. Reynolds, T.F. Quatieri and R.B. Dunn.: Speaker Verification Using Adapted Gaussian Mixture Models. In: Digital Signal Processing, Vol. 10, 19-41, 2000.
8. C. Sanderson.: Automatic Person Verification Using Speech and Face Information. PhD Thesis, Griffith University, Australia, August 2002 (revised Feburary 2003).
9. C. Sanderson and S. Bengio.: Robust Features for Frontal Face Authentication in Difficult Image Conditions. Proc. 4th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA), Guildford, UK, 2003.
10. M. Saban and C. Sanderson.: On Local Features for Face Verificaton. Technical Report IDIAP-RR 04-36, IDIAP, 2004. (www.idiap.ch)
11. M. A. Turk and A. P. Pentland.: Eigenfaces for recognition. In: Journal of Cognitive Neuroscience, 3(1), pp. 71–86, 1991.
12. W. Zhao, R. Chellappa, A. Rosenfeld, and P. J. Phillips.: Face Recognition: A Literature Survey. UMD CfAR Technical Report CAR-TR948, 2000.