ON COMBINING EVIDENCE FOR RELIABILITY ESTIMATION IN FACE VERIFICATION

Krzysztof Kryszczuk and Andrzej Drygajlo

Signal Processing Institute, Swiss Federal Institute of Technology Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
phone: + (41) 21 6934691, fax: + (41) 21 6937600, email: (krzysztof.kryszczuk, andrzej.drygajlo)@epfl.ch
web: scgwww.epfl.ch

ABSTRACT

Face verification is a difficult classification problem due to the fact that the appearance of a face can be altered by many extraneous factors, including head pose, illumination conditions, etc. A face verification system is likely to produce erroneous, unreliable decisions if there is a mismatch between the image acquisition conditions during the system training and the testing phases. We propose to detect and discard unreliable decisions based on the evidence originating from the classifier scores- and signal domains. We present a method of combining the reliability evidence, nested in a probabilistic framework that allows high level of flexibility in adding new evidence. Finally, we demonstrate on a standard evaluation database (Banca) how the proposed methodology helps in discarding unreliable decisions in a face verification system.

1. INTRODUCTION

Appearance-based face verification from two-dimensional images is a difficult classification problem due to fact that the intra-class variability is frequently greater than the separation between the class of genuine claims and the class of impostors. The appearance of an individual's face can be altered by a wide range of factors, ranging from pose, facial expression, and illumination variations, to the physical/optical characteristics and settings of the capture device. Numerous authors attempted dealing with the common adversities in the capture conditions that reduce the class separability. For instance, photometric normalization methods devised to cope with adverse illumination problems have been studied in great detail [6],[8],[10]. A lot of attention has been also paid to the problem of variable head pose [15]. Proposed methods help reduce the recognition errors to a greater or smaller extent. However, invariably they do not eliminate them.

Therefore there is a need for the estimation of decision uncertainty in the process of identity verification. The goal of the reliability estimation is to find out to what extent a verification decision can be trusted. This problem has recently received considerable attention and has been studied in the context of various biometric modalities [2],[3],[9],[14], including face verification [2],[3],[9]. The decision certainty estimate is frequently referred to as *confidence* [2],[12] or *reliability* [4],[5],[9],[14]. In accordance with the probabilistic formulation of decision reliability proposed in [14], a set of auxiliary information (evidence) is needed in order to arrive at an estimate of reliability. In [14], the classification scores and a single signal level quality measure (Signal to Noise Ratio) are used as evidence in a Bayesian network-based reliability estimator. In a similar fashion, a set of three signal-level quality measures is used to estimate the decision reliability in a face verification task [9].

In this paper we build on and extend the ideas presented in [9] and [14], and apply them in a face verification scenario. Specifically the contributions of this paper are as follows:

- Instead of using a Bayesian network wrap-up, we apply a simpler and more transparent Gaussian Mixture Model (GMM) approach for reliability estimation.
- We use score-derived and signal level quality measures together, concatenated into one evidence vector instead of separating classifier scores from quality measures in evidence modeling. We analyze how adding new evidence impacts the system performance.
- We interpret the verification process with reliability measures as two parallel classification problems: reliable/unreliable and accept/reject. The methodology bears resemblance to the Error/Reject tradeoff [4], but in our work we use separate sets of features for class separation and for reliability estimation. Unlike in [14] we do not attempt to correct decisions deemed unreliable since we do not have any grounds to perform such a correction.
- We analyze the influence of the reliability thresholding on relative accuracy gain of the classifier rather then setting it to a preset value of 0.5. In this way we explicitly show how collected evidence helps predicting erroneous decisions. At the same time, we present the reliability thresholding as a tradeoff between classifier accuracy and the relative volume of decisions labeled as unreliable.
- We propose a criterion for evaluating the overall performance of reliability estimation in a verification system.

This paper is structured as follows: Section 2 gives the details of the database and the classifier used, Section 3 presents the concept of classification reliability. Section 4 elaborates on the proposed quality measures used as the evidence in reliability estimation. Sections 5 and 6 provide a description of the proposed method of combining evidence for reliability estimation, followed by the experiment description and the discussion of the findings. Section 7 concludes this paper with a summary of presented results.

2. DATABASE AND CLASSIFIER DETAILS

We nested the experiments in a standard testing protocol for face verification - the P protocol defined for the Banca database (face part, English) [1]. In our work we used manually localized and geometrically normalized face images: the position of the eye centers are fixed. This constraint allowed us to eliminate the influences of imprecise face localization on the system errors, and hence to pinpoint the impact of image quality variation.

In our experiments we used a face verification scheme implemented in similar fashion as presented in [5]. Images from Banca database (English part) were used to build the world model (520 images, 26+10 individuals (g1 or g2 subsets, respectively), 384 components in the GMM). Client models were built using a recursive adaptation of the Gaussian component means from the world model, as described in [13]. The adaptation relevance parameter was set to 10, and the number of iterations was set to 3. The images used in the experiments were cropped, photometrically normalized by histogram equalization, and rescaled to the size of 64×80 pixels.

3. ESTIMATING DECISION UNCERTAINTY: THE CONCEPT OF RELIABILITY

In a face verification system, but as well in any other biometric authentication system, one can be interested, beside the actual classification decision (choice between two classes), in the degree of trust one can have that the classifier made a correct decision. This degree of trust is referred to as the reliability of the decision. The concept of reliability has been introduced already in [4] but its notion is rather intuitive than probabilistic. In [5] reliability parameters are derived from class posterior probabilities with no account for the signal quality. We adopt the probabilistic definition of the decision reliability *R*:

$$R = P\left(D_C \middle| E\right), \qquad (1)$$

where D_C denotes a correct classification decision and E denotes the supporting evidence [9]. The evidence may consist of information from the domains of classifier scores (score domain), features used by the classifier (feature domain), and the biometric presentation itself (signal domain).

Score domain evidence is what is used to estimate the reliability of the classification decision in the absence of any lower-level (feature or signal) information. As an example of this strategy one may consider the computation of posterior probabilities [2],[4]. However, classification scores may not be enough to accurately estimate the classification decision reliability in the presence of a mismatch between the conditions present during the acquisition of the biometric presentations (signals) used in the training and testing phases. An example how the condition mismatch can cause unreliable verification decisions in face verification is shown [7]. Reliability estimation is therefore essential in systems that may be affected by a condition mismatch. Reliability estimation is a process that is independent and parallel to the choice between an acceptance and rejection of the biometric presentation (Figure 1).

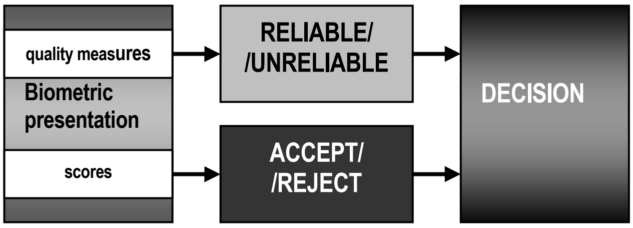

Figure 1: Biometric verification system with reliability measures.

Essentially, the reliability estimation turns unimodal biometric verification into two two-class classifiers (first reliable/unreliable, then accept/reject. Note that the term 'reject 'has been used in [4] to denote discarding of unreliable decisions. We use the term 'reject' as a claim rejection, in accordance with the terminology used in biometric verification). Following the probabilistic nature of the reliability estimation given by Equation 1, a decision of labeling a classification decision as reliable or unreliable depends on a chosen reliability threshold TR from the <0,1> range. In this framework, a reliability threshold of zero is equivalent to considering all decisions as reliable.

Decisions labeled as unreliable, depending on the architecture and purpose of the system, may be discarded and a new presentation may be requested [14], or the system may assume the 'safe state' [5], which in the case of biometric verification might be a rejection.

4. QUALITY MEASURES AS EVIDENCE FOR RELIABILITY ESTIMATION

It is difficult to define quantitatively the quality of a face image since there is no clear answer as to what features are essential for a successful face recognition. Given a cropped and geometrically normalized face image, a typical face verification system consists of the image preprocessing, feature extraction and classification stages (Figure 2). At each of those stages a quality assessment can be performed. We are interested in a *relative* quality measurement, taken in respect to the reference quality of the images used during system training. Such relative quality measures can be therefore treated as mismatch estimators.

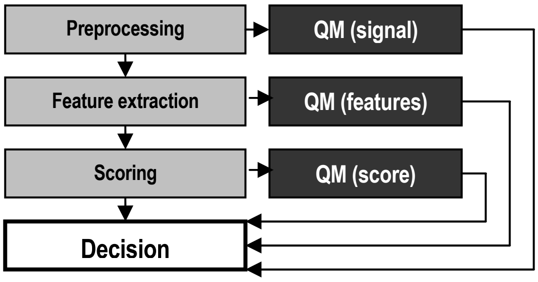

Figure 2: Stages of a face verification system with quality assessment

As Figure 2 shows, the information from low-level stage (signal level) flows up and impacts higher-level stage processing, including the decision-making stage. At the lowest, signal level, the exact impact of the image quality mismatch on the final decision is difficult to predict, but the quality degradation itself can be addressed directly. At the score level, the impact of the scores on the decisions is evident, but the sources of the impact are hard to trace. Hence the quality measures at each of the levels can be viewed as sources of complementary information about the verification process.

In this paper we discuss the use of signal- and score-level quality measures for face verification, since the use of those measures is universal for any classifier that allows a direct access to the classification scores (before thresholding). The use of feature-level quality measures is classifier-specific and therefore out of the direct scope of this paper.

4.1 Score-level: absolute distance between the log-likelihood ratio and the decision threshold

The distance between the decision threshold and the actual value of the log-likelihood ratio (score) is a measure of how insensitive the decision is to the departure from an optimal threshold value. The actual sign of the distance, while of crucial importance for the verification decision, is immaterial in the reliability estimation. We define the quality measure *QM_1*:

$$QM_1 = \left| L(X \mid \lambda_C) - L(X \mid \lambda_W) - \Theta_D \right|, \qquad (2)$$

where Θ_D is the classification threshold optimized on the development set [10], λ_C is a probabilistic *client model* and λ_W is the *world model*, a model that represents a generic, client-independent distribution of features [10]

4.2 Score-level: Sum of log-likelihoods

The goal of the likelihood ratio-based verification is to find if the feature vector is better represented by λ_C or by λ_W. Log-likelihood ratio does not help detecting a situation when neither of the models represents the data adequately (in the presence of a condition mismatch). We propose to compute a measure of the match of the input image with either of the two models, or both simultaneously. For given feature set X originating from the image I we define the quality measure *QM_2*:

$$QM_2 = L(X \mid \lambda_C) + L(X \mid \lambda_W) . \qquad (3)$$

Since $L(X|\lambda_C)$ and $L(X|\lambda_W)$ are expressed in the log-domain, Equation 3 is mathematically equivalent to a multiplication of likelihoods. The model λ_C should represent a subset of faces modeled by λ_W since a face of a particular individual is an instance of the generic class of faces. Therefore very low values of *QM_2* correspond to images that are well accounted for by neither $L(X|\lambda_C)$, nor $L(X|\lambda_W)$.

4.3 Signal-level: Correlation with an average face template

The goal of the relative quality measurement is to determine to what degree the quality of the testing image departs from that of the training images, which can be achieved by creating an *average face template*. An average face template is built out of all the face images whose quality is considered as reference. We have built an average face template using PCA reconstruction, in similar fashion as described in [16]. Specifically, we used the first eight averaged eigenfaces to build the template. Two average face templates built of images from the Banca database are found in Figure 3.

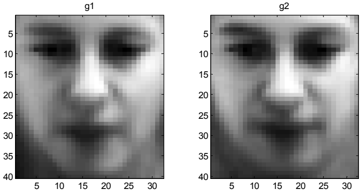

Figure 3: Average face templates ***AVF*** built using training images defined in the Banca P protocol, for the datasets g1 and g2.

For the experiments presented in this paper we have created two average face templates from the training images prescribed by the P protocol (clients from the groups g1 and g2). It is noteworthy that the average face templates created from the images of two disjoint sets of individuals are strikingly similar. It is also apparent that high-resolution details are lost, while low-frequency features, such as head pose and illumination, are preserved.

Therefore, in order to obtain a measure *QM_3* of similarity of low-frequency face images, we propose to calculate the Pearson's cross-correlation coefficient between the face image *I*, whose quality is under assessment, and the respective average face template *AVF*:

$$QM_3 = corrcoeff\left(AVF, I\right) \qquad (4)$$

4.4 Signal-level: Image sharpness estimation

The cross-correlation with an average image gives an estimate of the quality deterioration in the low-frequency features. At the same time that measure ignores any quality deterioration in the upper range of spatial frequencies. The absence of high-frequency image details can be described as the loss of image sharpness. In the case of the Banca database, the images collected in the *degraded* conditions suffer from a significant loss of sharpness. In order to estimate the sharpness of an image I of $x \times y$ pixels, we compute *QM_4*, the mean of intensity differences between adjacent pixels, taken in both the vertical and horizontal directions:

$$QM_4 = \frac{1}{2}\left[\frac{1}{(x-1)y} \sum_{m=1}^{y} \sum_{n=1}^{x-1} \left| p_{n,m} - p_{n+1,m} \right| + \right. \\ \left. + \frac{1}{(y-1)x} \sum_{m=1}^{y-1} \sum_{n=1}^{x} \left| p_{n,m} - p_{n,m+1} \right| \right] . \qquad (5)$$

5. COMBINING EVIDENCE AND ERROR PREDICTION

In order to adhere to the P evaluation protocol defined for the Banca database, we have built a model of the quality measures using the development set, and applied it to predict unreliable classifier decisions on the testing set. For each dataset (g1 and g2), we have constructed two concurrent probabilistic models of the quality measure distributions: one for the correct, and one for the erroneous classifier decisions on the development dataset. We refer to those models as λ_{DC} and λ_{DF}, respectively. The models are built as follows: for each testing image *I* from the development set we construct a vector of *n* quality measurements V_{QM}:

$$V_{QM} = \left(QM_1, QM_2, \dots, QM_n\right) \tag{6}$$

The vectors are separated into those for which the classifier decision was correct (D_C), and erroneous (D_F). We build GMM-based models of the distribution of $V_{QM}|D_C$ and $V_{QM}|D_F$ (λ_{DC} and λ_{DF}):

$$\begin{aligned} \lambda_{DC} &= \{\mu_{DC}, \sigma_{DC}, \alpha_{DC}\} \equiv p\left(V_{QM}\middle|D_C\right) \\ \lambda_{DF} &= \{\mu_{DF}, \sigma_{DF}, \alpha_{DF}\} \equiv p\left(V_{QM}\middle|D_F\right)' \end{aligned} \tag{7}$$

where μ, σ *and* α are the parameter vectors of the mixture of Gaussians. In our work we used the Expectation-Maximization algorithm to train the models. We assumed the statistical conditional independence of QM_1, QM_2 and QM_3, and therefore chose to build the models with diagonal covariance matrices. We used 12 Gaussian components per mixture.

Consequently, we used the models trained for the dataset g1 to estimate the reliability of classifier decisions obtained using the dataset g2, and vice-versa. For each testing image we computed conditional log-likelihoods $L(V_{QM}|\lambda_{DC})$ and $L(V_{QM}|\lambda_{DF})$. The decision reliability estimate assuming equal prior probabilities of encountering a reliable and unreliable decision, following Equation 1 and the Bayes' rule, is then given by:

$$R = P\left(D_C\middle|V_{QM}\right) = \frac{L\left(V_{QM}\middle|\lambda_{DC}\right)}{L\left(V_{QM}\middle|\lambda_{DC}\right) + L\left(V_{QM}\middle|\lambda_{DF}\right)}. \tag{8}$$

6. EXPERIMENTAL RESULTS

We have conducted a set of experiments in which we have computed the reliability estimates using Equation 8 for all test presentations from group g1 and g2, using different combinations of evidence: $E_1 = \{QM_1\}$, $E_2 = \{QM_1, QM_2\}$, $E_3 = \{QM_1, QM_2, QM_3\}$, $E_4 = \{QM_1, QM_2, QM_3, QM_4\}$.

After having estimated the reliability of each decision, the obtained value was compared with the reliability threshold T_R which represents how much we are willing to trust the classifier. If the estimated reliability falls below the preset threshold, the decision is classified as unreliable and discarded.

In order to evaluate the prediction accuracy of proposed models we have checked what the accuracy of the classifier was after the decisions labeled as unreliable had been discarded. We have been changing the reliability decision threshold $T_R \in \langle 0,0.95 \rangle$ in 0.05 increments and computing the accuracy of the classifier after having discarded unreliable decisions. The situation when T_R=0 corresponds to a system without any reliability estimators: all decisions are equally and fully trusted.

In order to better evaluate the performance gain the results are presented in relative terms: accuracy gain is expressed as the percentage of the accuracy of the classifier without reliability measures (or T_R=0). The relative gain of accuracy A_D is plotted against the reliability threshold T_R in Figure 4.

Forcing higher classification accuracy comes at a cost of having to discard a number of unreliable decisions [4], therefore the accuracy gain cannot be used alone as the performance evaluator. In general, it is desirable to make as many reliable decisions as possible, at highest possible accuracy. We hence present the number of remaining, reliable decisions N_D as the function of T_R (Figure 5). N_D is also expressed in relative terms as the percentage of the total count of test decisions taken (2730 decisions for g1 and 2730 decisions for g2).

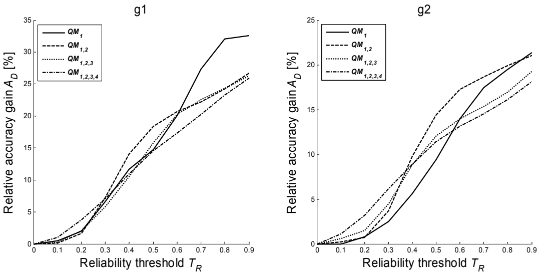

Figure 4: Relative accuracy gain $\boldsymbol{A_D}$

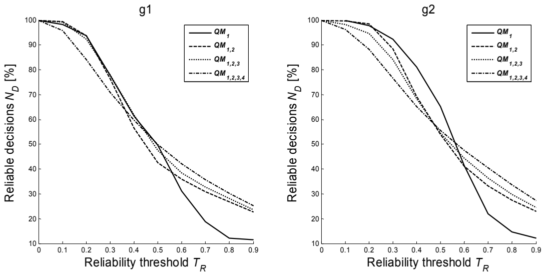

Figure 5: Relative count of reliable decisions $\boldsymbol{N_D}$

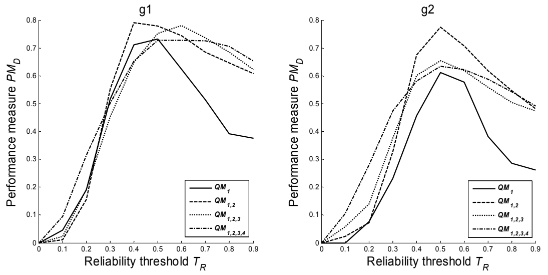

Figure 6: Combined performance estimate $\boldsymbol{PM_D}$

When optimizing a verification system with reliability measures we wish to maximize both the accuracy gain and the number of remaining, reliable decisions after reliability-thresholding. This can be achieved by maximizing the product PM_D=(A_D·N_D). The obtained values of PM_D can be interpreted as an overall performance measure for given reliability threshold T_R. This measure is shown in Figure 6.

6.1 Discussion of the results

Experimental results presented in Figures 4, 5 and 6 demonstrate that proposed method of combining evidence for reliability estimation allows for error prediction in the scenario of a face verification system. All curves that represent relative accuracy gain A_D in Figure 4 are monotonically growing with the reliability threshold T_R, which proves that achieved accuracy gain is not happening by chance. This result means that the more certainty in decision making is desired, the more accurate the classifier indeed is, after the decisions deemed unreliable have been discarded. At the same time, the relative amount of discarded decisions increases monotonically with T_R. The proposed performance measure PM_D is a formalization of a trade-off between desired reliability of the decision made, and the amount of the decisions discarded. The graphs shown in Figure 6 allow for an easy comparison between reliability estimation using different pieces of evidence, and can as well be used to compare different reliability estimators. Also, graphs presented in Figures 4, 5 and 6 make it possible to adjust the value of the reliability threshold to fit the performance requirements of a particular application.

Results presented above also show that a careful choice of evidence is important for achieving desired system properties. While all evidence combinations do bring a gain in classification accuracy, using only QM_1 as evidence offers best accuracy gains. At the same time, however, the amount of decisions labeled as unreliable grows considerably. Adding QM_2 to the evidence vector slightly reduces the accuracy gain for the highest values of T_R, while increasing the relative accuracy gain for the remaining values of T_R and significantly decreasing the number of discarded decisions for T_R>0.6. Adding the signal quality measures QM_3 and QM_4 increases the number of remaining decisions N_D but does not bring further accuracy improvement.

Figure 6 shows that the performance measure PM_D yields maximal values for the reliability threshold close to 0.5. This finding agrees with an intuitive understanding of the concept of reliability, according to which R<0.5 could be attributed to decisions taken by random.

7. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a method of combining evidence for reliability estimation in a face verification system. We have demonstrated that proposed method allows for an easy concatenation of evidence originating from different domains (signal, score). We analyzed how different evidence combinations impact the error prediction in a reference face verification scenario (Banca). Finally, we have proposed a combined performance measure of reliability estimation.

In ongoing research we are extending the proposed framework to multiple classifier and multimodal biometric verification, in particular in order to address the problem of discarded biometric presentations labeled as unreliable.

REFERENCES

[1] E. Bailly-Baillire, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Poree, B. Ruiz, J.-Ph. Thiran, "The BANCA Database and Evaluation Protocol", *Proc. 4th AVBPA*, Guilford, UK, 2003

[2] S. Bengio, C. Marcel, S. Marcel and J. Mariethoz, "Confidence Measures for Multimodal Identity Verification", *In: Information Fusion*, Vol. 3, No. 4, pp. 267-276, 2002.

[3] J. Bigun, J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, “Multimodal biometric authentication using quality signals in mobile communications”, *Proc. 12th International Conference on Image Analysis and Processing*, Mantova, Italy, 2003.

[4] C.K. Chow, “On Optimum Recognition Error and Rejection Tradeoff”, *IEEE Transactions on Information Theory*, Vol. 16, No.1, pp. 41-46, January 1970.

[5] L.P. Cordella, P. Foggia, C. Sansone, F. Tortorella, M. Vento, “Optimizing the Error/Reject Tradeoff for a Multi-Expert System using the Bayesian Combining Rule”, Proc. SSPR/SPR, pp: 716 – 725, 1998.

[6] R. Gross and V. Brajovic, “An Image Preprocessing Algorithm for Illumination Invariant Face Recognition”, *Proc. 4th AVBPA*, Guilford, UK, 2003.

[7] K. Kryszczuk and A Drygajlo, “Addressing the vulnerabilities of likelihood-ratio-based face verification”, *Proc. 5th AVBPA*, Rye Brook NY, USA 2005.

[8] K. Kryszczuk and A Drygajlo, “Gradient-based image segmentation for face recognition robust to directional illumination”, *Proc. Visual Communication and Image Processing*, Beijing, China, 2005.

[9] K. Kryszczuk, J. Richiardi, P. Prodanov, A. Drygajlo, “Error Handling In Multimodal Biometric Systems Using Reliability Measures”, *13th European Signal Processing Conference (EUSIPCO 2005)*, Antalya, Turkey, 2005.

[10] S. Lucey and T. Chen, “A GMM Parts Based Face Representation for Improved Verification through Relevance Adaptation”. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Vol. 2, Washington, USA, 2004.

[11] 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 ICBA*, Hong Kong, 2004.

[12] N. Poh, S. Bengio, “Improving Fusion with Margin-Derived Confidence In Biometric Authentication Tasks”, In: *Proc. AVBPA 2005*, Rye Brook NY, USA, 2005.

[13] D.A. Reynolds, T.F. Quatieri and R.B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models”. *Digital Signal Processing*, Vol. 10, pp. 19-41, 2000.

[14] J. Richiardi, P. Prodanov, and A. Drygajlo, “A probabilistic measure of modality reliability in speaker verification,” In: *Proc. of the ICASSP 2005*, Philadelphia, USA, 2005.

[15] C. Sanderson, S. Bengio and Y. Gao,” On Transforming Statistical Models for Non-Frontal Face Verification”, Appears in: Pattern Recognition, Vol. 39, No. 2, 2006, pp. 288-302, http://dx.doi.org/10.1016/j.patcog.2005.07.001

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