On Face Image Quality Measures

Krzysztof Kryszczuk and Andrzej Drygajlo
Swiss Federal Institute of Technology (EPFL)
krzysztof.kryszczuk@epfl.ch, andrzej.drygajlo@epfl.ch

Abstract

The classification reliability is an essential problem in biometric verification systems. In the presence of a possible mismatch between system's training and testing conditions, a measure of that mismatch is necessary in order to estimate the degree of trust one can have in the classification decisions. In the case of face verification from plain images, there are many factors that can affect the image quality, and thus create a condition mismatch. In this paper we identify the most commonly encountered causes of face image degradation due to the recording conditions, and propose corresponding quality measure methods. We show on publicly available databases that proposed methods can be effectively used to estimate the verification reliability using a probabilistic framework.

1. Introduction

In the recent years a growing interest in face recognition systems is evident. There exist advanced projects of large-scale deployment of biometric recognition systems in important public establishments, border crossing checkpoints and large sport events venues. Due to the ease of data collection and public acceptability factors, face image verification is the modality of choice in many applications. At the same time however, face verification systems have not quite yet reached the required maturity for their large-scale deployment. In particular, reported error rates are typically higher than those of other biometric modalities (fingerprints, iris). Commonly encountered error rates can be compared by examining reported experimental results [10,12,13]. 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 legitimate identity claimant class 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 variation, 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 [5,7,9]. A lot of attention has been also paid to the problem of variable head pose and facial expression. Proposed methods help reduce the total recognition errors to a greater or smaller extent. However, invariably they do not eliminate them.

Hence there is a need to estimate the decision uncertainty in the process of identity verification. The goal of the reliability estimation is to decide 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 [3,4,8,13], including face verification [3,4,8]. The decision certainty estimate is frequently referred to as *confidence* [3,11] or *reliability* [8,13]. Many of those methods call for quality measures of the biometric samples used in the verification process, and it has been shown that quality measures can improve the performance of a biometric verification system [4,8,11]. To our knowledge, no systematic analysis of the problem of face image quality assessment has yet been presented. We hence propose a set of automatic face quality measures and a probabilistic framework of their use that can be used in any face verification system.

Estimation of the quality of face images is a non-trivial problem because of a multitude of behavioral and extraneous conditions that can simultaneously affect the face appearance in the image. We have previously proposed a set of automatic face image quality measures to estimate the effects of non-frontal illumination [7] and additive noise [6], but their application in the estimation of the decision uncertainty has not been sufficiently elicited.

In this paper we present a systematic analysis of the problem of face image quality estimation. We analyze the possible sources of degradation of the face image quality, and their impact on the image itself. We also consider the effects of quality degradation on the scores used by the classifier in order to arrive at the verification decisions. We propose new face image quality estimators devised to cope with images of degraded illumination, contrast and sharpness, and with variable face pose. Finally, we demonstrate on the example of the Banca database [2] and a local DCT feature-based classification system [9,10] how the discussed face quality measures can be used to assess the decision reliability in a face verification system.

This paper is structured as follows: Section 2 defines the concept of reliability and justifies the need for measuring the quality of face images. Section 3 focuses on the possible sources of face image degradation. Sections 4 and 5 give the details of the database, experimental protocols and verification system used in presented work. Section 6 elaborates on the proposed quality measures, followed by the experiment description and the discussion of the findings in Section 7. Section 8 concludes this paper with a summary of presented results.

2. 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 have that the classifier made a correct decision. This degree of trust is referred to as the *reliability* of the decision. In general, decision reliability R is defined as a conditional probability:

$$R = P(DC|E), \tag{1}$$

where DC denotes a correct classification decision and E denotes the supporting *evidence* [8]. 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).

The score domain evidence 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. However, score domain information 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 is shown in [1] (speaker verification) and [6] (face verification).

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).

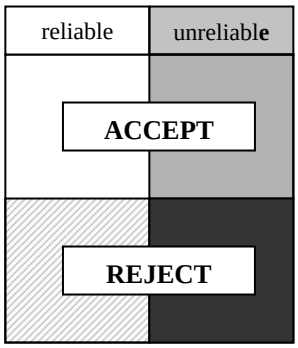

Figure 1: Classification and reliability estimation.

Essentially, the reliability estimation turns a two-class classification problem into a three-class problem (*accept*, *reject* and *unreliable*) or a cascade of two two-class classifiers (first *reliable/unreliable*, then *accept/reject*). Following the probabilistic nature of the reliability estimation assumed in Equation 1, a decision on labeling a classification decision as reliable or unreliable depends on a chosen reliability threshold T_R 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 [13], or the system may assume the 'safe state', which in the case of biometric verification might be a rejection.

3. Possible sources of face image degradation

The possible sources of face image degradation may come, among other factors, from:

- **illumination variation,**
- **additive noise,**
- **head pose and facial expression,**
- **image sharpness and geometric distortions caused by the imaging optics.**

The changes of face illumination can dramatically alter its appearance. Particularly strong changes result from the use of directional light sources that may cause heavy self-shadowing of the face as well as specular reflections. In [7] we have presented an automatic method of estimating the degradation of the image due to adverse directional illumination based on face segmentation using local gradient variance.

Additive noise frequently appears in low-light image acquisition conditions and is a function of the quality of signal amplification in the imaging sensor. We have proposed an automatic method of quality estimation for face images contaminated with additive noise based on two-dimensional correlation with an average face template [6,14]. In this paper we propose to extend the application of the correlation-based quality estimation by applying it to the testing images in which the head pose and facial expression as well as the illumination, differs from those from the training gallery.

The image sharpness has a deciding influence on the level of fine details in the face image. Decline in the image sharpness is frequently caused by the use of a low-quality optical imaging system (e.g. low-end webcams), but also by improper functioning of automatic focusing systems or inappropriate settings of a fixed focus camera. In this paper we propose an automatic method of image sharpness estimation based on average intensity differences between neighboring pixels.

Geometric distortions caused by the imperfections of the imaging systems are very difficult to estimate without the use of special pattern templates. Typically they also do not change during the operation of the imaging system. Therefore they are out of scope of the work presented in this paper.

For our experiments presented in this paper we have used the Banca database (English part) [2], since it is one of the reference databases [10], and since it contains images collected in different recording conditions, and using various imaging devices.

4. The Banca database and evaluation protocols

The face part of the English Banca database consists of still face images of 52 individuals. The images were captured in *controlled*, *degraded* and *adverse* conditions. For each of the conditions, 4 separate recording sessions were organized. For the details on the Banca database the reader is referred to [2]. An example of the images collected under *controlled*, *degraded* and *adverse* conditions can be found in Figure 2.

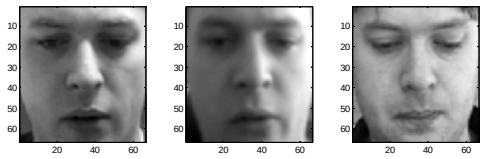

Figure 2: Example of the images collected in the *controlled, degraded* and *adverse* scenarios (left to right).

In our experiments, we have adhered to the P protocol. The P protocol assumes that the *controlled* conditions are the reference for training, and testing is performed on images collected in all conditions.

As the reader can gather from the sample shown in Figure 2, the images collected in the *degraded* and *adverse* conditions differ qualitatively from those collected in *controlled* conditions. In particular, the imaging device used varies from one condition to another: therefore the image sharpness is not constant. This difference is particularly pronounced for the *degraded* conditions. In respect to the *controlled* conditions there exists a noticeable difference in the illumination conditions and the head pose (the subjects are mostly looking down) in the *adverse* conditions.

In our work we use manually localized and geometrically normalized face images: the position of the eye centers are fixed. This constraint allows us to eliminate the influences of erroneous face localization on the system errors, and hence to pinpoint the impact of image quality variation. The problem of face quality measurement is not irrelevant to automatic face localization: changes in image quality are likely to affect the face detection accuracy.

5. The DCTmod2-GMM face verification

In our experiments we have used a face verification scheme implemented in similar fashion as presented in [9]. The images from Banca database (English part) were used to build the world model (520 images, 26+10 individuals (g1 or g2 subsets, respectively), 384 Gaussians in the mixture). Client models were built using a recursive adaptation of the Gaussian component means from the world model, as described in [12]. 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.

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 log-likelihood ratio:

$$LLR(X) = L(X \mid \lambda_W) - L(X \mid \lambda_C), \quad (2)$$

where $L(X|\lambda_C)$ and $L(X|\lambda_W)$ are the log-likelihoods of the set of vectors X given λ_C (the model of client C), and λ_W (the *world model*). The value of $LLR(X)$ is compared to a threshold Θ, whose value has been optimized to minimize the half-total error rate (HTER) on the development set (in accordance with the Banca protocol P). The average HTER were comparable to the state-of the art algorithms [10].

6. Quality measures for face images

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 scoring stages (Figure 3).

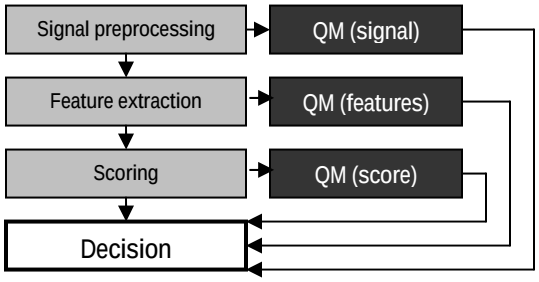

Figure 3: Stages of a face verification system with quality assessment

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.

As Figure 3 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 quality change 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. Therefore the quality measures from 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, because 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.

6.1. Signal-level: Correlation with an average face image

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 modeled 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 [14]. Specifically, we have 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 4.

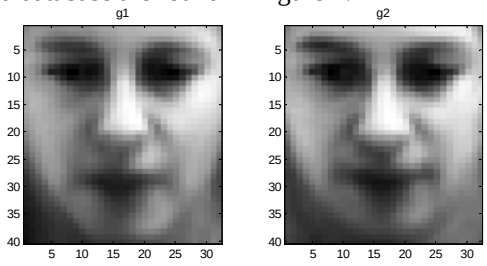

Figure 4: Average face template built using training images defined in the Banca P protocol, for the datasets g1 and g2, respectively.

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 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.

$$QM_1 = corrcoeff\left(AVF, I\right) \quad (3)$$

6.2. 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. An example of this deterioration can be found in Figure 2.
In order to estimate the sharpness of an image I of $x \times y$ pixels, we compute the mean of intensity differences between adjacent pixels, taken in both the vertical and horizontal directions:

$$QM_2 = \frac{1}{2}\left[\frac{1}{(x-1)y}\sum_{m=1}^{y}\sum_{n=1}^{x-1}\left|p_{n,m} - p_{n+1,m}\right| + \frac{1}{(y-1)x}\sum_{m=1}^{y-1}\sum_{n=1}^{x}\left|p_{n,m} - p_{n,m+1}\right|\right]. \quad (4)$$

6.3. Score-level: Sum of log-likelihoods

The concept of likelihood ratio-based verification, as expressed by Equation 2, is to establish 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 (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:

$$QM_3 = L(X \mid \lambda_C) + L(X \mid \lambda_W). \quad (5)$$

Since $L(X|\lambda_C)$ and $L(X|\lambda_W)$ are expressed in the log-domain, Equation 5 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 correspond to images that are well accounted for by neither $L(X|\lambda_C)$, nor $L(X|\lambda_W)$.

7. Experiments

7.1. Error distributions across recording sessions

The Banca database consists of images collected in three distinct recording conditions, organized in 12 different sessions. We assume the experimental conditions and thus the image quality to be constant within one session. The testing according to the P protocol consists of 2×2730 verification decisions on the subsets g1 and g2. Figure 5 shows the distribution of the images from all 12 sessions in the testing set (both g1 and g2).

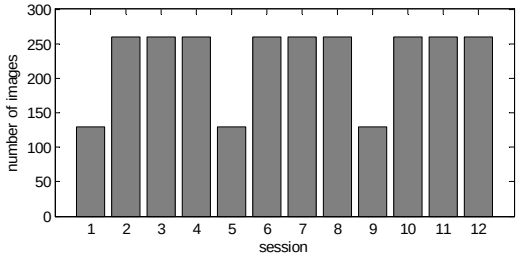

Figure 5: Distribution of images from 12 Banca sessions in the testing process, P protocol. The distribution is identical for datasets g1 and g2.

The error rate distributions for all 12 sessions in terms of HTER are shown in Figure 6. Each plot bar represents the error rate for the given session.

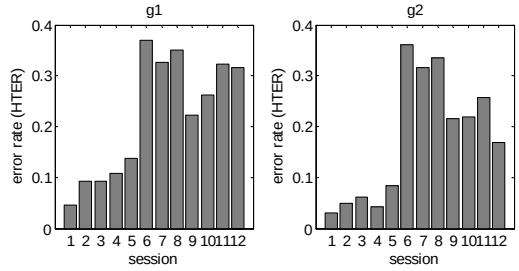

Figure 6: Error rates for each of the 12 recording sessions of the Banca database. Testing according to the P protocol, errors reported in terms of HTER, separately for datasets g1 and g2.

It is evident from the graphs in Figure 6 that the error rates are not evenly distributed among the sessions and experimental conditions. The bulk of errors are concentrated around sessions collected in the *degraded* (sessions 5-8) conditions, fewer errors are present in the *adverse* (sessions 9-12) condition data, and finally lowest error rates, understandably, are recorded for the *controlled* conditions (sessions 1-4). The differences in error rates from condition to condition and from

session to session can be attributed to the mismatch in the recording conditions, hence to the degradation of the relative quality of the testing images with respect to the training data.

7.1. Distributions of quality measures

A good quality measure should model and predict well a degradation of the system performance (increased error rates) due to a particular factor that impacts the image quality. In order to find how the quality measures proposed in Section 6 are suitable for predicting recognition errors, we calculate the quality measures QM_1, QM_2 and QM_3 for every test image. Mean values of the quality measures over all experimental sessions are shown in Figures 7, 8 and 9. In those figures, each plot bar represents the mean of a corresponding quality measure for the given session.

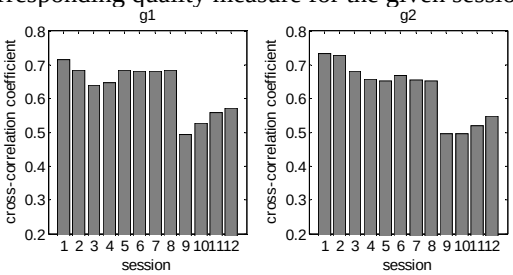

Figure 7: distribution of QM_1 means for each recording session.

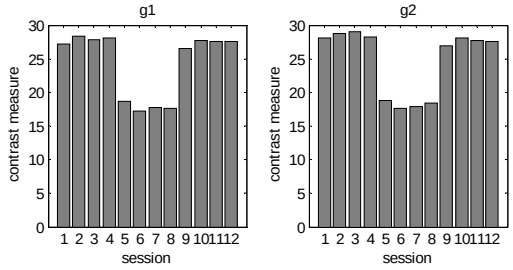

Figure 8: distribution of QM_2 means for each recording session.

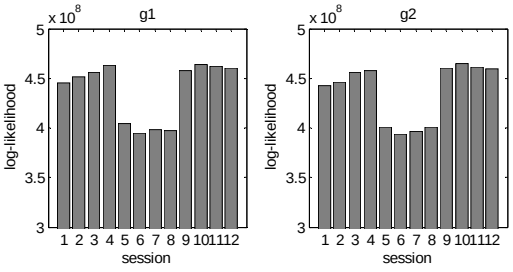

Figure 9: distribution of QM_3 means for each recording session.

As the Figures 7,8 and 9 show, none of the quality measures alone predicts well the classifier errors, since each of them is responsible for a certain aspect of the complex concept of face image quality. Distribution of the means of QM_1 respond strongly to the departure from the reference illumination conditions and pose variation, present in the *adverse* conditions (sessions 9 – 12). Since the average face template is itself a low-frequency model, the change in image sharpness between the reference and the *degraded* conditions passes mostly unnoticed when applying QM_1.

Complementary to QM_1, QM_2 is designed to respond to the sharpness change in the testing images. As Figure 8 shows, QM_2 plays this role well, sporting a clear dip in the average quality measure scores for the session recorded in the *degraded* conditions (5 – 8).

QM_3 is a score-domain quality measure and its mean values reflect not only the quality of the image itself, but it is also affected by the feature robustness and the goodness of fit of the models λ_C and λ_W. Figure 9 shows that the corresponding quality measure scores according to QM_3 are higher for the *adverse* conditions, and lower for the *degraded* conditions, in respect to the reference images (session 1). A higher score for QM_3 should not be interpreted as a 'better fit'. It actually means that the test images fit both λ_C and λ_W, hence the decision of acceptance or rejection has a small error margin [11].

7.2. Modeling of the quality measures and error prediction

In order to adhere to the P evaluation protocol defined for the Banca database, we have decided to build a model of the quality measures using the development set, and apply 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_n from the development set we construct a vector of quality measurements V_{QM}:

$$V_{QM}^{n} = \left(QM_1^n, QM_2^n, QM_3^n\right). \quad (6)$$

The vectors are separated into those corresponding to the correct (*DC*), and erroneous (*DF*), classifier decisions. We build the GMM-based models of the distribution of $V_{QM}|DC$ and $V_{QM}|DF$ (λ_{DC} and λ_{DF}):

$$\lambda_{DC} = \{\mu_{DC}, \sigma_{DC}, \alpha_{DC}\}$$
$$\lambda_{DF} = \{\mu_{DF}, \sigma_{DF}, \alpha_{DF}\}, \quad (7)$$

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, following Equation 1 and the Bayes' rule, is then given by:

$$R = P(DC|V_{QM}) = \frac{\gamma \cdot L(V_{QM}|\lambda_{DC})}{L(V_{QM}|\lambda_{DC}) + L(V_{QM}|\lambda_{DF})}, \quad (8)$$

where γ is a constant reflecting the ratio of priors. In our experiments we assumed equal priors, hence γ=1. Figure 10 shows the results of the mean reliability (error prediction) estimation for datasets g1 and g2. For easier comparison with Figure 6 we present the prediction results in terms of error probabilities (1-*R*).

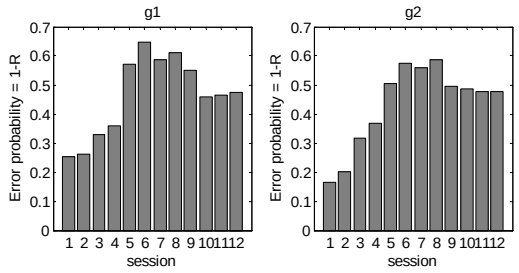

Figure 10: Mean reliability of verification per session. Prediction presented in the terms of error probabilities (1-*R*).

7.3. Evaluation of the error prediction accuracy

After having estimated the reliability of a decision, it is necessary to compare the obtained value to a preset 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 (Figure 1) 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 results of this evaluation can be found in Figure 11:

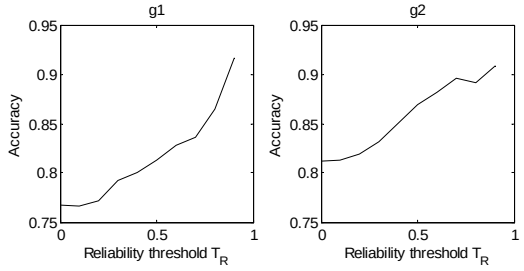

Figure 11: Accuracy of the classifier after having discarded unreliable decisions, as the function of the reliability threshold T_R.

The situation when T_R=0 corresponds to a system without any reliability estimators: all decisions are equally and fully trusted.

Forcing higher classification accuracy comes at a cost of having to deal with a number of unreliable decisions. The number of the decisions discarded depends to a large degree on the available data itself. For the P protocol of the Banca database, the percentages of discarded decisions in the function of chosen reliability threshold T_R are shown in Figure 12.

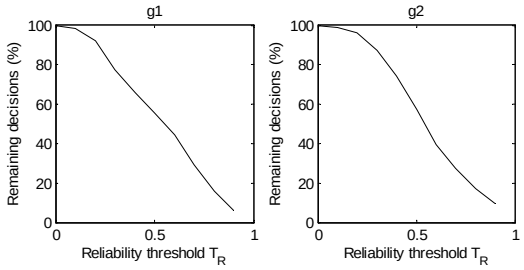

Figure 12: Percentage of remaining decisions after reliability-based thresholding, as a function of the reliability threshold T_R.

7.4. Discussion of the experimental results

The mean predicted errors per session presented in Figure 10 follow the tendencies of the actual verification errors, shown in Figure 6. However, it can be noticed that the predicted error rates are generally higher than the actual ones. This rather careful estimation of reliability can be explained by the fact

that in our experiment only the quality measures were used to predict the system performance. We wish to bring to the reader's attention the fact that we have not used the actual log-likelihood ratio, the basis of the verification decisions, in our reliability measurements. It can be expected that including the log-likelihood ratio as one of the quality measures would improve the predictive power of the proposed methods. It is worth noticing that any number of new quality measures can be added to the quality measure vector (Section 7.2) without the need to modify the structure of the reliability estimator – with the exception of the need to re-train the models λ_{DC} and λ_{DF} in order to include the new dimensions.

Prediction evaluation results presented in Figure 11 prove that proposed face quality measures can be effectively used to estimate the face verification reliability. Verification accuracy that grows monotonously with the reliability threshold T_R proves that erroneous decisions are effectively discarded based on the obtained reliability measure. In order to further improve the system performance, a sequential repair strategy can be applied [13].

The error prediction for the experimental session 5 seems to be particularly inadequate, as the bar graph in Figure 10 shows. This indicates the proposed set of quality measures fails to capture certain face image features that are important in the process of face verification. Future work is planned in order to address this problem.

A small local decline in accuracy on the curve for dataset g2, in Figure 11, is caused by the fact that as the number of remaining decisions gets smaller, certain individuals, whose face, although of high quality, is inherently difficult to distinguish from the others ('sheep/wolves').

8. Conclusions

We have presented a novel concept of reliability in face verification. We have justified the need to perform a set of quality measurements of face images in order to estimate the verification reliability. We have performed an analysis of the factors impacting the quality of a face image and affecting the reliability of face verification. We have proposed a set of quality measures for face images operating at the signal and score level, and have shown a probabilistic scheme of reliability assessment using the proposed measures. Finally, we have demonstrated on the example of the Banca database that the proposed system can be effectively applied to assess the decision reliability in face verification.

10. References

[1] M. Arcienega, A. Alexander, P. Zimmermann, A. Drygajlo, "A Bayesian network approach combining pitch and spectral envelope features to reduce channel mismatch in speaker verification and forensic speaker recognition", In: *INTERSPEECH-2005*, Lisbon, Portugal 2005.

[2] 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*, Surrey, UK, 2003

[3] 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.

[4] 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.

[5] R. Gross and V. Brajovic, "An Image Preprocessing Algorithm for Illumination Invariant Face Recognition", *Proc. 4th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, Guilford, UK, 2003.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

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

[13] 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.

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