ERROR HANDLING IN MULTIMODAL BIOMETRIC SYSTEMS USING RELIABILITY MEASURES

Krzysztof Kryszczuk, Jonas Richiardi, Plamen Prodanov, Andrzej Drygajlo

Swiss Federal Institute of Technology Lausanne (EPFL)
1015 Lausanne, Switzerland
phone: + (41) 021 693 4691, fax: + (41) 693 7600,
email: {krzysztof.kryszczuk, jonas.richiardi, plamen.prodanov, andrzej.drygajlo}@epfl.ch
web: http://scgwww.epfl.ch

ABSTRACT

In this paper, we present a framework for predicting and correcting classification decision errors based on modality reliability measures in a multimodal biometric system. In our experiments we use face and speech experts based on a recently proposed framework which uses Bayesian networks. The expert decisions and the accompanying information on their reliability are combined in a decision module that produces the final verification decision. The proposed system is consequently shown to yield higher decision accuracy than the corresponding unimodal systems.

1. INTRODUCTION

Biometric verification systems that use a single biometric modality often have to contend with adverse environmental conditions such as background noise in speaker verification and illumination changes in face-based verification. Attempting to improve the performance of unimodal biometric systems in such situations may not prove to be effective because of these inherent problems. Therefore, combining multiple independent modalities which are not degraded by the same environmental effects will afford robustness to adverse conditions. Multimodal biometric identity verification has frequently been shown to outperform unimodal approaches. Many fusion schemes have been proposed for combining multiple classifiers [1].

Common fusion methods include some form of a priori judgement of the average reliability of the decisions of each of the unimodal classifiers, typically based on performance over a development set [2],[1]. This average modality reliability information can be applied to weight the unimodal classifier decisions during the fusion process.

The drawback of this approach is that it does not take into account the fact that individual decisions depend on the acquisition condition of the data presented to the expert as much as they depend on the discriminating skills of the classifier.

Recently, signal quality and impostor/client score distributions have been used to train weights information for classifier combination in multimodal biometric verification [3]. The quality measures are used during the training of the decision module, and do not play an explicit role in the rectification of unimodal decisions before the fusion. The quality measures for particular modalities are subjective quality tags which are manually assigned to the training and test data.

A method for estimating the reliability of the individual classifier decisions that can be used in rectifying erroneous decisions was proposed in [4]. The method uses a Bayesian network trained to predict classification errors given the classification score, classifier decision, and automatically obtained auxiliary information about the quality of the biometric data presented to the unimodal classifier. A system using a speech expert consisting of a speech classifier combined with a decision reliability estimator was shown to significantly reduce the total classification error rates for speech-based biometric verification. In the unimodal scenario, an unreliable verification decision entails a request for a repeated presentation. In the presence of a second biometric trait available, such a sequential repair strategy can be replaced by a parallel one, where the unreliable decision of one unimodal classifier can be replaced by a more reliable decision for another modality.

In this paper, we present an embodiment of this parallel multimodal repair strategy, using speech and face experts and a multimodal fusion module. The proposed method yields higher accuracy in prediction and correction of the verification decisions than each of the unimodal experts alone.
This paper is structured as follows: in Section 2 we present the framework of the estimation of the verification decision reliability using Bayesian networks, Section 3 describes the multimodal database used in our work, Sections 5 and 4 discuss the quality measures used by the face and speech experts, respectively. Section 6 treats of the multimodal decision combination module. Experimental results are shown in Section 7 and their discussion accompanied by the conclusions are found in Section 8.

2. VERIFICATION DECISION RELIABILITY ESTIMATION WITH BAYESIAN NETWORKS

We define decision reliability for a given modality *MR* as the probability that the classifier for this modality has taken a correct verification decision given the available evidence, i.e. the probability $P(MR|E)$. The evidence E that provides information about the state of *MR* can be selected from several levels: signal domain, feature domain, score domain, or decision domain itself. In the present work, for each modality we use a vector of signal-domain quality measures $\mathbf{QM}$, classifier score information *Sc* and classifier decision *CID* ($CID = 1$ if the classifier thinks the biometric presentation belongs to the claimed client, otherwise $CID = 0$). Furthermore, in training a decision reliability estimator, it is crucial to provide the ground truth about the user TID ($TID = 1$ if the biometric presentation really belongs to the claimed client, otherwise $TID = 0$) so that the influence of the event "the user is a client" on other variables can be taken into account in modelling. These sources of information and

their interrelations are modelled probabilistically using the Bayesian Network shown on Fig. 1. For more details on the rationale behind the creation of this model, the reader is referred to [4].

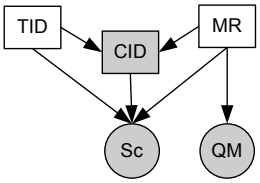

Figure 1: Bayesian network for modality decision reliability estimation

The Bayesian network is used for providing values for $P(MR|E)$, which in our case is $P(MR|CID,Sc,\mathbf{QM})$. Inference on $P(MR|CID,Sc,\mathbf{QM})$ is only possible once the conditional distribution parameters for the variables have been learned from training examples. The network parameters can be estimated using a maximum likelihood (ML) training technique [5]. Figure 2 provides an overview of a modality expert consisting of the baseline classifier for a modality and the corresponding Bayesian network estimating the decision reliability . The classifier part of the expert is trained from clean held-out data which is not used again (see Section 7). The reliability estimator is trained on sets of variable values $(CID,Sc,\mathbf{QM},TID)$ obtained by feeding degraded conditions biometric data to the classifier and the environmental conditions measurer. The environmental conditions measurer provides values for the $\mathbf{QM}$ variable as described in Sections 5 and 4.

It should be noted that TID is only observed during training.

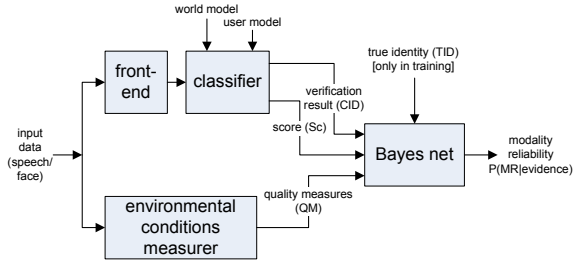

Figure 2: Modality expert with modality classifier and modality reliability estimator

Probabilistic decision reliability for each modality, e. g. for speech $P(MR_s = 1|CID,Sc,\mathbf{QM})$ and for face $P(MR_f = 1|CID,Sc,\mathbf{QM})$ can be used to enhance the accuracy of the final decision of the multimodal verification system.

3. DATABASE AND EXPERIMENTAL CONDITIONS

For the purpose of the experiment a multimodal database of chimeral users was created using face images from the YaleB database [6] and speech samples from the BANCA database [7]. The multimodal database used in our experiments was created using data for all 10 users from YaleB database and 10 selected users from the BANCA database. The consistency of the assignment of face and speech data to users was preserved throughout the entire experiment.

3.1 Face modality data

The choice of the YaleB database for face images was dictated by the fact that to the best of our knowledge it is the only face image database available that offers face images in fully controlled clean and degraded illumination conditions. The limitation of YaleB database is its size: it contains face images of only 10 users. The face part of the resulting database consisted of data for 10 users, 11 recording conditions, 9 presentations per user and condition. The recording conditions included clean (non-degraded) set of images, used for face classifier training, and 10 sets of images recorded in the presence of illumination coming from various angles. Face images were cropped out manually.

3.2 Speech modality data

The BANCA database was chosen because it provides a large amount of training data per user: 2 files per session (about 20 sec. each) x 2 microphones x 12 sessions. In our case we used only the data of the first 10 users from microphone 2, which has a much larger dynamic range. The first 4 sessions are "clean" conditions, the next 4 sessions are "degraded" conditions, and the last 4 sessions are "adverse" conditions. The files have clicks and a spectral line at 16 kHz, but only the first file of session 1 for each user was pre-processed to remove the clicks in the middle and at the end of the file. The rest was left untouched because our interest was in testing a continuum of acoustic conditions.

4. SPEAKER VERIFICATION AND QUALITY MEASURES

The speech-based classifier is trained by segmenting the first file of session 1 for each user into 4 segments of approximately 5 seconds. 12 Mel-Frequency Cepstral Coefficients with first and second order time derivatives are extracted with cepstral mean normalisation. The features are modelled by a Gaussian Mixture Model of 64 Gaussian components with diagonal covariance matrices. Log-likelihood ratio scores are produced using a 64 Gaussians world model (trained from the pooled training data of all users) for normalisation. The thresholds are trained a priori. This classifier provides the *CID* and *Sc* variables to the reliability estimator.

The signal-to-noise ratio (SNR) contains information about the level of acoustic noise in the signal, which is one of the main factors of signal quality degradation. Thus, the quality measure used for speech is an SNR-related measure. The SNR can be defined as the ratio of the average energy of the speech signal divided by the average energy of the acoustic noise in dB. We perform speech/pause segmentation using an algorithm based on the "Murphy algorithm" described in [8]. We then assume that the average energy of pauses is associated with that of noise. Our SNR-related quality measure (SQM) is given by the formula:

$$SQM = 10\log_{10}\frac{\sum_{i=1}^{N} Is(i)s^2(i)}{\sum_{i=1}^{N} In(i)s^2(i)} \qquad (1)$$

where $\{s(i)\}, i = 1,\ldots,N$ is the acquired speech signal containing N samples, $Is(i)$ and $In(i)$ are the indicator functions of the current sample $s(i)$ being speech or noise during pauses (e.g. $Is(i)$=1 if $s(i)$ is a speech sample, $Is(i)$=0 otherwise).

5. FACE VERIFICATION AND QUALITY MEASURES

The baseline face classifier was based on DCTmod2 features and a GMM classifier, built in an identical fashion as described in [9]. The world model was trained using all images from BANCA database, French part, controlled condition, to assure that the face verification results are not specific to the database used for verification. The thresholds are trained a posteriori using equal error-rate criterion. As in the case of speech, we define quality measures to quantify the difference of the signal with nominal conditions. Three quality measures are combined to create a face quality measure vector $\mathbf{FQM} = [FQM_1, FQM_2, FQM_3]'$.

FQM_1 is computed by comparing the mean pixel intensity value of the normalized test image I to the mean $_T$ of pixel intensity values of the normalized face images from the training set. The quality measure FQM_1 is the distance between the means.

$$FQM_1 = \frac{1}{X \cdot Y} \sum_{x=1}^{X} \sum_{y=1}^{Y} I_{x,y} - \ _T, \tag{2}$$

where X, Y are the dimensions of I.

FQM_2 is computed as follows: a 2-dimensional normalized cross-correlation between the test image and the average face template T_F is calculated. The average face template is built using the images from the training set using PCA reconstruction [10].

$$FQM_2 = max[C_{norm}(I, T_F)], \tag{3}$$

where $C_{norm}(I, T_F)$ denotes the normalized 2-dimensional cross-correlation between the test image I and the average face template T_F [11].

FQM_3 is computed as follows: the images from the training set are divided into $N \times M$ blocks of 8×8 pixels with 4 pixels horizontal and vertical overlap. For each block $b_{x,y}$, the variance of the pixel intensity values $_p$ is calculated. For a vector of pixel variance values originating from the corresponding block $b_{x,y}$ of all images from the training set, a mean $_{x,y}$ and variance $_{x,y}$ is computed. In the quality estimation process, the test image is divided into blocks in identical way as for training images. Again, for each block, pixel intensity variance $_{x,y}$ is calculated and the likelihood FQM_3 is found:

$$FQM_3 = \sum_{x=1}^{N} \sum_{y=1}^{M} \ln\left(\frac{1}{_{x,y}\sqrt{2}} e^{\frac{_{x,y} - \ _{x,y}}{2 \ _{x,y}^2}} \right) \tag{4}$$

6. MULTIMODAL DECISION FUSION WITH RELIABILITY INFORMATION

Figure 3 presents the schematic diagram of the system used in our experiment. Biometric data of an individual (face image and speech) are corrupted by extraneous conditions: in the case of speech additive noise, and in the case of the face illuminance difference from the nominal lighting. The speech and face acquisition process consists of all the signal-domain preprocessing and normalisation steps ([9], [4]) that make the speech data and face image usable for the experts (see Figure 2). Each of the experts accepts two inputs: the conditioned data from the acquisition process and the identity claim. On the output, the experts produce the verification decisions CID_f and CID_s (for face and speech , accordingly) and modality reliability information MR_f and MR_s, on the base of which the multimodal decision module (see Table 1) produces the final verification decision.

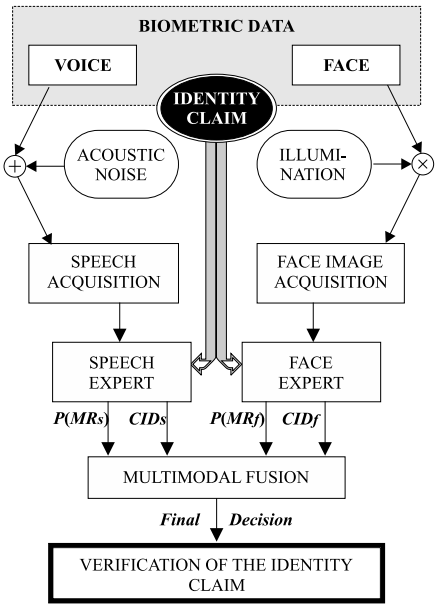

Figure 3: Multimodal biometric verification system with reliability information

The fusion of the verification information coming from face and speech experts is performed using the classifier decisions and the modality reliability data. If both experts agree on the decision, the decision is preserved. If they are in disagreement, the decision is taken in accordance to Table 1.

Face	Speech	Final decision
$CID_f = 1$	$CID_s = 1$	1
$CID_f = 1$	$CID_s = 0$	1: $if\ P(MR_f = 1) > P(MR_s = 1)$, 0: *otherwise*
$CID_f = 0$	$CID_s = 1$	1: $if P(MR_f = 1) < P(MR_s = 1)$, 0: *otherwise*
$CID_f = 0$	$CID_s = 0$	0

Table 1: Decision table for multimodal decision module

7. EXPERIMENTS AND RESULTS

The experiment to evaluate the performance of the proposed system was designed as follows. Out of the total pool of classification results for face and speech a balanced sample containing equal count of correct accepts, correct rejections, false acceptances and false rejections was selected. Two thirds of the volume of this sample was used in training the Bayesian nets of face and speech experts, and this portion of data was not used for testing. The remaining part of the sample volume was used to test the system. For each combination of genuine and impostor claims present in the sample, the face expert decision CID_f, its reliability measure $P(MR_f = 1)$, the speech expert decision CID_s and its reliability measure $P(MR_s = 1)$ were fed into the multimodal

decision module. The final decision of the module was compared to the a priori known ground truth data. Due to the initial balancing of the data, the reference accuracy of the face and speech classifiers (without classifier error prediction and repair) is by definition 50%. The results presented in this section are reported in reference to this value. Because we assumed equal costs for false accepts and false rejects, the overall accuracy is computed by averaging the accuracy for clients and the accuracy for impostors. The experiment was repeated 100 times, giving a total of 159805 final classification decisions. The mean accuracies over all client and impostor accesses are presented in Table 2.

	Face	Speech	Combined
baseline mean	50.0%	50.0%	n/a
mean	75.6%	74.6%	90.1%
std	2.82%	1.2%	0.9%
improvement	25.6%	24.6%	40.1%[1]

Table 2: Experiment results: mean accuracy of the final identity verification decisions for each modality and their combination with 100 cross-validation passes

As described in Section 6, the final decision could be unanimous, or be made upon the comparison of the modality reliability information in the case of disagreement. Table 3 shows the statistics of the decisions for the 100 cross-validation experiments.

	Face wins	Speech wins	Unanimous
mean	163 (10.2%)	293 (18.3%)	1142 (71.5%)
std	14 (0.9%)	20 (1.3%)	44 (1.2%)

Table 3: Agreement statistics

The relative performance improvement is a predictor of the performance improvement of a system that assigns equal cost to false accepts as false rejects, because of the way the experiment was designed. It can be expected that in such a scenario the multimodal authentication system will be 40.1% more accurate than each of the baseline unimodal verification systems alone.

8. DISCUSSION AND CONCLUSIONS

We have presented a framework for a multimodal classification system using Bayesian networks for modelling decision reliability measures for each modality classifier. The reliability measures are explicitly involved in the final multimodal decision rule to resolve disagreement between the classifiers in favour of the more reliable modality. Within this framework, we introduced the use of automatic signal-domain quality measures which play an important role in the rectification of unimodal classifier errors. We have applied the above framework to a biometric verification system working with face and speech data. We have shown that the majority of erroneous decisions of unimodal classifiers can be rectified by the use of the decision reliability measures. The results of the reported experiments show that, depending on the application and system utilization scenario, the overall accuracy of the system can be significantly improved. It is worth noticing that under current multimodal decision scheme (Section 6) the system is forced to take the decision of the expert who reports higher reliability, even if this reliability for both experts may be very low (data from both modalities unreliable). This is a potential source of verification errors. In our future work we intend to address this deficiency by applying more sophisticated decision scheme, e.g. a sequential parallel repair scenario.

[1] this denotes the improvement over any of the unimodal classifiers that do not use reliability information. Improvement over the unimodal experts that use modality-specific reliability information is 14.5% for face and 15.5% for speech.

REFERENCES

[1] A. Ross, A. Jain, and J.-Z. Qian, "Information fusion in biometrics," in *Proc. of 3rd Int'l Conference on Audio- and Video-Based Person Authentication (AVBPA)*, Sweden, 2001, pp. 354–359.

[2] R. Roli, J. Kittler, G. Fumera, and D. Muntoni, "An experimental comparison of classifier fusion rules for multimodal personal identity verification systems," in *Proc. Third International Workshop on Multiple Classifier Systems*, June 2002, pp. 325–226.

[3] J. Bigun, J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Multimodal biometric authentication using quality signals in mobile communications," in *Proc. 12th international conference on image analysis and processing*, Mantova, Italy, Sept. 2003, pp. 2–11.

[4] J. Richiardi, P. Prodanov, and A. Drygajlo, "A probabilistic measure of modality reliability in speaker verification," in *Proc. IEEE ICASSP 2005*, 2005.

[5] K. Murphy, *Dynamic Bayesian networks: representation, inference and learning*, Ph.D. thesis, U.C. Berkeley, July 2002.

[6] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[7] S. Bengio, F. Bimbot, J. Mariethoz, V. Popovici, F. Poree, E. Bailly-Balliere, G. Matas, and B. Ruiz, "Experimental protocol on the banca database," Technical Report IDIAP-RR 02-05, IDIAP, 2002.

[8] D. Reynolds, *A Gaussian mixture modeling approach to text-independent speaker identification*, Ph.D. thesis, Georgia Institute of Technology, 1992.

[9] C. Sanderson and S. Bengio, "Robust features for frontal face authentication in difficult image conditions," in *Proc. 4th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, Guildford, UK, 2003.

[10] K. Kryszczuk and A. Drygajlo, "Color correction for face detection based on human visual perception metaphor," in *Proc. of the Workshop on Multimodal User Authentica-tion*, Santa Barbara, USA, 2003, pp. 138–143.

[11] Robert M. Haralick and Linda G. Shapiro, *Computer and Robot Vision*, vol. 2, Addison-Wesley, 1992.