Multimodal Authentication using Qualitative Support Vector Machines

F. Alsaade, A. Ariyaeeinia, L. Meng and A. Malegaonkar

University of Hertfordshire, Hatfield, Herts, AL10 9AB, U.K.

{F.Alsaade, A.M. Ariyaeeinina, L.Meng, A.Malegaonkar}@herts.ac.uk

Abstract

This paper proposes an approach to enhancing the accuracy of multimodal biometrics in uncontrolled environments. Variation in operating conditions results in mismatch between the training and test material, and thereby affects the biometric authentication performance regardless of this being unimodal or multimodal. The paper proposes a technique to reduce the effects of such variations in multimodal fusion. The proposed technique is based on estimating the quality aspect of the test scores and then passing these aspects into the Support Vector Machine either as features or weights. Since the fusion process is based on the learning classifier of Support Vector Machine, the technique is termed Support Vector Machine with Quality Measurement (SVM-QM). The experimental investigation is conducted using face and speech modalities. The results clearly show the benefits gained from learning the quality aspects of the biometric data used for authentication.

Index Terms: Multimodal biometric authentication, score level fusion, quality measurement, support vector machine

1. Introduction

The automatic verification of the identities of individuals is becoming an increasingly important requirement in a variety of applications, especially, those involving automatic access controls. Examples of such applications are teleshopping, telebanking, physical access control, and the withdrawal of money from automatic telling machines (ATMs). The identification means used traditionally in this context, such as passwords and personal identification numbers (PINs), can easily be compromised or forgotten.

It is widely recognised that through the use of biometrics, the above-mentioned problems with the conventional identification methods can be avoided. Due to the potential advantages offered by biometrics, this has been the subject of significant research in recent years. Despite this, however, there are still limitations associated with the use of a single biometrics type for the purpose of authentication. Some important causes of such limitations with unimodal biometrics are non-universality and ease of spoofing. The former refers to occasion on which a user is incapable of providing biometric data due, for example, to some form of disability. These limitations can be overcome, to a large extent, through the use of multimodal biometrics. With such an approach, spoofing would be considerably difficult as it has to be carried out simultaneously with more than one type of biometrics. Potentially, it also provides the possibility of using different combinations of biometrics for individuals and thereby enhances the universality of the technology as a whole.

However, a main attraction of multimodal biometrics, which is the subject of this paper, is that it provides the possibility of increasing the authentication accuracy beyond what is obtainable using a single biometrics type. On the other hand, a possible drawback of multimodal biometrics is the complexity of the architecture of the authentication system. This in turn may reduce the computational speed of the authentication process.

In general, multimodal biometrics is based on the understanding that features extracted from different modalities possess complementary information about an individual's identity [1]. Consequently, combining the information, obtained from different modalities, should be more useful than that extracted from any single modality involved. The information combination can be carried out at various levels. Examples these are the feature level, score level and decision level. It has, however, been reported that the most appropriate and effective approach to multimodal biometrics is through the fusion of information at the score level [2]. One of the issues of concerns in such fusion is the intrinsic variation in scores due to external factors. For example, speaker-dependant and speaker-independent variations in the speech signal can affect the quality of the score obtained in the trial from a legitimate speaker. Background lighting conditions and inter-session variations in sensor characteristics can also affect the score obtained from modalities such as face, and lips. These factors can also adversely affect the outcome of multimodal biometrics, if a fixed fusion procedure developed based on clean training data, is deployed.

To tackle this problem, it would be logical to consider the relative levels of contamination in different biometric data used in the fusion process. To date, various adaptive approaches have been proposed for this purpose [3]. A main aspect of these techniques is the quantification of the quality aspect of the test trial in the scoring process, and fusing this aspect with the actual score. This is typically based on estimating some reference parameters from the development scores and then estimating the quality aspect for the test scores using these parameters.

The main aim of this work is to investigate whether such quality aspects can be beneficial in the score level fusion especially with those classifiers that have good learning mechanism such as Support Vector Machines (SVMs) [4]. Hence this study is mainly focussed on SVM based score level fusion. For this purpose, two schemes are proposed in this work. These schemes mainly differ in the manner of passing the quality related information to SVM. The paper provides details of these schemes and experimentally evaluates their effectiveness in multi-modal fusion.

The rest of the paper is organised as follows. Section 2 provides details of the proposed schemes. Section 3 describes the experimental investigation and discusses the results, and Section 4 gives the overall conclusions.

2. The Proposed Scheme

In this technique the quality aspect of testing samples is quantified and then passed into a SVM. This process involves estimating the quality of the development data by measuring some parameters for the development data and then incorporating these parameters in the quality estimation of the test scores. This quantification is similar to that described in [3] and is described as follows.

In the case of a two-class problem (Clients / Impostors), let $M(f/s)$ be the development scores for face or speech, (where (f/s) is used to denote that a measure is applied to either face or speech modality) and let the client and impostor scores from each modality be given as

$$C_{M(f/s)} \cong \left\{\mu^{C}_{M(f/s)}, \sigma^{C}_{M(f/s)}\right\} \tag{1}$$

$$I_{M(f/s)} \cong \left\{\mu^{I}_{M(f/s)}, \sigma^{I}_{M(f/s)}\right\} \tag{2}$$

where $\mu^{C}_{M(f/s)}$ and $\sigma^{C}_{M(f/s)}$ are the mean and variance for the client scores from each modality - face or speech, $\mu^{I}_{M(f/s)}$ and $\sigma^{I}_{M(f/s)}$ are the mean and variance for the impostor scores from each modality - face or speech.

The quality of samples of a modality (face or speech in this paper) is determined by the characteristics of the scores obtained with the development and test samples of that modality. The quality of the face scores (Q_f) and speech scores (Q_s) are calculated as follows:

$$Q_{(f/s)} = D_{M(f/s)} \times T_{E(f/s)} \tag{3}$$

where $Q_{(f/s)}$ is the quality for face or speech, D is the quality of the development data, T is the quality of the test data and $E(f/s)$ is the subset of scores from the test data which is used to determine the quality of the test data.

Based on equation (3), the computation for the quality of samples is divided into two steps.

2.1. Estimation of the quality aspects for the development data samples

$D_{M(f/s)}$ in equation (3) denotes the quality of the development data for face or speech scores. It is computed based on the scores obtained in the development phase as follows.

$$D_{M(f)} = \frac{l_{M(s)}}{l_{M(s)} + l_{M(f)}} \,, \tag{4}$$

$$D_{M(s)} = \frac{l_{M(f)}}{l_{M(s)} + l_{M(f)}} \,, \tag{5}$$

where $l_{M(f)}, l_{M(s)}$ are computed during the development phase using equation (6).

$$l_{M(f/s)} = \sqrt{\frac{\left(\sigma^{C}_{M(f/s)}\right)^2}{N^{C}_{M(f/s)}} + \frac{\left(\sigma^{I}_{M(f/s)}\right)^2}{N^{I}_{M(f/s)}}} \,, \tag{6}$$

where $N^{C}_{M(f/s)}$ is the total number of clients in the development data for each modality - face or speech, and $N^{I}_{M(f/s)}$ is the total number of impostors in the development data for each modality - face or speech.

2.2. Estimation of the quality aspects for the test data samples

$T_{E(f/s)}$ in equation (3) represents the quality of the test data for face or speech scores. These quality aspects are calculated using a subset of the test data as follows

$$T_{E(f)} = \frac{k_{E(s)}}{k_{E(s)} + k_{E(f)}} \,, \tag{7}$$

$$T_{E(s)} = \frac{k_{E(f)}}{k_{E(s)} + k_{E(f)}} \,, \tag{8}$$

where $k_{E(f/s)}$ is computed during the test phase as follows

$$k_{E(f/s)} = \frac{\left| \dfrac{(E(f/s) - \mu^{C}_{M(f/s)})^2}{\sigma^{C}_{M(f/s)}} - \dfrac{(E(f/s) - \mu^{I}_{M(f/s)})^2}{\sigma^{I}_{M(f/s)}} \right|}{\mu^{C}_{M(f/s)}} \,. \tag{9}$$

The quality measurements for face scores Q_f and speech scores Q_s are passed to a SVM using two different approaches. These two approaches and the motivation behind them are discussed in the next section.

2.3. Methods of passing the quality aspects to SVM

In this work, two approaches for passing the quality of the test scores to SVM have been studied. The first approach is based on passing quality aspects in the individual modality as a separate feature for SVM. In the second approach, quality aspects in each of the modalities are fused with the respective scores and then the combined scores are passed as a feature to support vector machines. These approaches are described in the following subsections.

2.3.1. Quality aspects as independent features (Q-IF)

In this approach, SVM is fed with four input vectors, two of these present the development data (M_f, M_s) and the other two present the quality of both the development and test data (Q_f, Q_s), as shown in Figure 1.

In this approach, during development stage the quality of the face and speech scores is passed into the SVM as new features alongside the development scores. The SVM uses these four input vectors to tune its parameters to fit the test data. In the test stage, four input vectors are passed into the classifier, with two of them presenting the quality of the test data. These are computed based on the parameters obtained from the development data. The other two vectors present the test data itself.

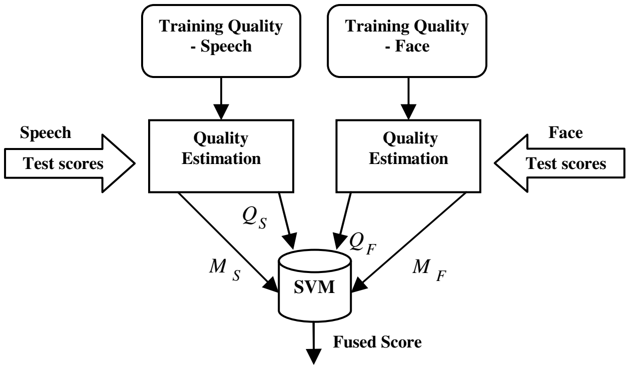

Figure 1. *Proposed Scheme of SVM-QM using quality aspect as separate features*

2.3.2 Modality specific fusion of quality aspects(Q-MSF)

This approach is shown in Figure 2.

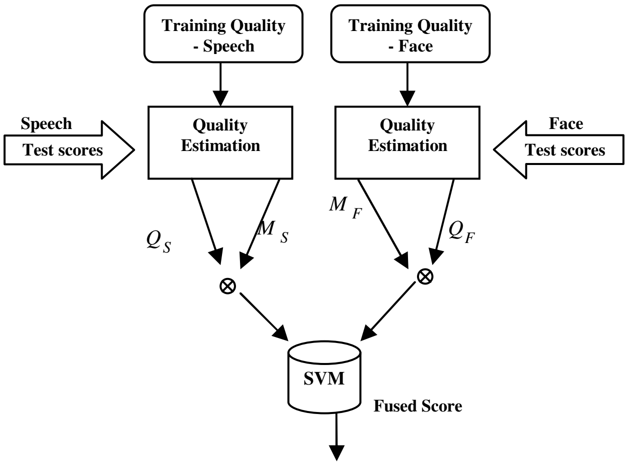

Figure 2. *Proposed Scheme of SVM-QM using quality aspect as weights*

In this approach, quality for face and speech scores is considered as weights. These weights must be in the interval of 0 and 1 with the condition of $\sum Q_{(f,s)} = 1$. To achieve this, the quality obtained from equation (3) is normalized using the following two steps.

$$S_i = Q_{f_i} + Q_{s_i} \qquad (10)$$

where S_i is the summation of the i^{th} face quality Q_{f_i} and its corresponding i^{th} speech quality Q_{s_i}. The weight for the i^{th} face or speech scores $W_{(f/s)_i}$ is obtained as,

$$W_{(f/s)_i} = \frac{Q_{(f/s)_i}}{S_i} \qquad (11)$$

These weights for face or speech scores are then multiplied by their corresponding face or speech scores, respectively. The results of these multiplications, two weighted input vectors, are then passed on to SVM.

In the development phase, the two weighted input vectors are used in order to optimise the parameters of SVM. The parameters obtained in the development phase are then used in the test phase to classify the test scores according to SVM.

3. Experimental Investigation

3.1. The database

Experiments are conducted using a subset of the XM2VTS database [6]. This is a multi-session database containing synchronized image and speech data obtained from 295 subjects, recorded during four sessions which are taken at one month intervals [6]. For the purpose of this study, the subjects in the database are divided into three sets, the training set is used to train client models, the development set is used to obtain various parameters in the proposed scheme and the test set is used to investigate the performance. The training set consists of 200 client subjects, the development set consists of 25 non-client subjects and the test set consists of 70 non-client subjects. The total number of 200 client tests and 40000 non-client tests is used from the development data while the total number of client and non-client tests used in finding the quality of the test data is 200 and 40000 respectively. The rest of the test set, 200 client tests and 72000 non-client tests, is used to investigate the performance for the proposed scheme. This division is based on the framework of Lausanne protocol which is further described in [6].

3.2 Feature and Classifiers

This study is based on using one feature for the face modality and three different types of features for the speech modality. The face feature is Discrete Cosine Transform for Big Image (DCTb). On the other hand, the speech features are Linear Frequency Cepstral Coefficients (LFCC), Phase Auto Correlation (PAC) and spectral Sub-band Centroids (SSC).
In this study, all of the features (face/speech) are represented by Gaussian Mixture Model (GMM) classifiers. More information about the classifiers and features can be found in [6].

3.3. Testing with Fusion

The testing procedure involves combining the scores obtained from the face feature and one speech feature, using SVM classifiers. This constitutes 3 different combinations of features for the fusion purpose. Each of the score streams is normalised using Zero-score normalisation [5] before the fusion process. The fusion process is carried out using linear SVM [4]. The tests are conducted with and without learning the quality aspect of the test data.

3.4. Results and Discussions

In this study, the results obtained for the authentication tests are given in terms of Equal Error Rate (%EER) with 95% confidence interval. Table 1 shows the baseline results obtained using the individual features. It can be seen that, amongst these, the best EER is obtained with the speech feature of LFCC.

Feature		% EER ± CI 95
face	DCTb	1.87 ± 0.32
speech	LFCC	1.06 ± 0.18
	PAC	6.56 ± 1.07
	SSC	4.53 ± 0.76

Table 1. *Baseline results for uni-modal authentication.*

The results for the fusion exercise with and without learning the quality are given in Table 2, again in terms of Equal Error Rate (%EER) with 95% confidence interval.

	Feature types		SVM (without QM)	SVM-QM	
	Face	voice		Q-IF	Q-MSF
1	DCTb	LFCC	0.65±0.11	0.38±0.10	0.22±0.06
2		PAC	1.22±0.21	0.35±0.09	0.43±0.12
3		SSC	1.05±0.18	0.36±0.10	0.35±0.09

Table 2. *Bi-modal authentication with and without quality learning.*

It can be observed that the best results without quality learning process are obtained by combining the scores obtained from DCTb and LFCC feature. It can also be observed that the reduction in EER obtained by learning the quality of the data is quite significant. The lowest EER (0.22%) is observed in the case of DCTb-LFCC combination with SVM-QM. Such a result is observed when the quality is passed to the SVM as weights. The EER reduction in this case is 66% compared with the best result obtained without quality learning.

These results clearly show that learning the quality information of a score is useful for improving the performance of the multimodal authentication systems. A direct comparison of the EERs obtained using fusion with and without quality learning, together with the baseline EERs for each of the two cases of DCTb and LFCC is given in Figure 3 as a DET (Detection Error Trade-off) plot.

4. Conclusions

It can be concluded from this study that the combination of complementary information from the face and speech can improve the performance over single-modality. Amongst the two fusion schemes considered (SVM and SVM-QM), SVM-QM scheme has appeared to provide better performance in terms of reducing error rates. Such results prove that Linear SVM can benefit from the quality of the testing data in order to decrease the system error rates. Encouraging initial results of the proposed approaches motivate further research in order to exploit quality of the testing data in the fusion stage of multimodal biometric verification systems.

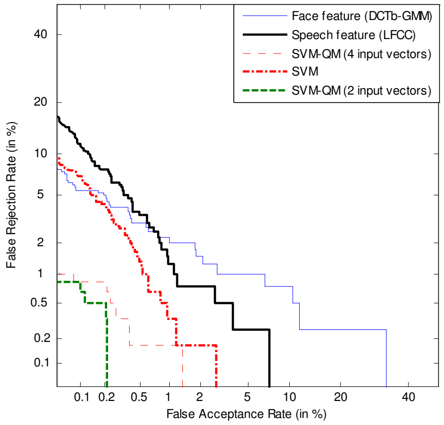

Figure 3. *DET plot for SVM fusion method with and without quality learning in uni-modal and bi-modal fusion.*

5. References

[1] Roli, F., Kitler, J., Fumera, G., and Muntoni, D. “An Experimental Comparison of Classifier Fusion Rules for Multimodal Personal Identity Verification Systems”, *Proc. Multiple Classifier Systems, Springer-Verlag*, 2002, pp. 325-336.

[2] Jain, Anil K., and Ross, A “Multibiometric systems” Communication of the ACM, v.47 n.1, January 2004.

[3] Sanderson, C., and Paliwal, K.K. “Information Fusion and Person Verification Using Speech and Face Information”, *IDIAP-RR* 02-33, 2003.

[4] Burges, C.J.C., “A tutorial on support vector machines for pattern recognition”. *Data Mining and Knowledge Discovery*, 2(2), pp. 955-974, 1998.

[5] Alsaade, F., Malegaonkar, A., and Ariyaeeinia A., “Fusion of Cross Stream Information in Speaker Verification”, *Proc. COST 275 Workshop on Biometrics on the Internet,* Hatfield, UK, pp. 63-66, Oct 2005.

[6] Poh, N., and Bengio, S., “Database, Protocol and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication”, in *Fifth Int'l. Conf. Audio- and Video-Based Biometric Person Authentication AVBPA*, 2005.