A Discriminative Parametric Approach to Video-based Score-level Fusion for Biometric Authentication

Norman Poh and Josef Kittler
CVSSP, FEPS, University of Surrey, GU2 7XH, UK
{n.poh,j.kittler}@surrey.ac.uk

Fuad Alkoot
HITN, PAEET, Alsalmia, Kuwait
f_alkoot@yahoo.com

Abstract

Video-based biometric systems are becoming feasible thanks to advancement in both algorithms and computation platforms. Such systems have many advantages: improved robustness to spoof attack, performance gain thanks to variance reduction, and increased data quality/resolution, among others. We investigate a discriminative video-based score-level fusion mechanism, which enables an existing biometric system to further harness the riches of temporarily sampled biometric data using a set of distribution descriptors. Our approach shows that higher order moments of the video scores contain discriminative information. To our best knowledge, this is the first time this higher order moment is reported to be effective in the score-level fusion literature. Experimental results based on face and speech unimodal systems, as well as multimodal fusion, show that our proposal can improve the performance over that of the standard fixed rule fusion strategies by as much as 50%.

1 Introduction

Thanks to the advancement of sensing technology as well as advancement in highly efficient computational algorithms and hardware computation platforms, it is now possible to acquire and process video biometric data in real time. There are at least three potential advantages in exploiting the temporal information for biometric person recognition: (i) as evidence of biometric liveness, for instance, using continuously acquired fingerprint images [1] or in audio-visual biometrics [2]; (ii) improved accuracy via variance reduction [8], and (iii) as a means to derive super-resolution biometric samples [14].

In this paper, we extend the concept of multi-sample score-level fusion to include the notion of time, which we refer to as short-term temporal fusion or simply temporal fusion. The aim here is to investigate if the overall system accuracy can exceed that of the conventional approach using such simple rules as mean, maximum or minimum of the scores. Rather than using these simple statistics, for each stream of data, we propose to coarsely characterise the distribution of the scores by a vector of distribution descriptors consisting of mean, standard deviation, skewness, median, the 25-th and 75-th percentiles, as well as minimum and maximum of the scores. The importance of these parameters with respect to the classification task is directly learnt from the data via logistic regression.

In essence, our proposal compares two score distributions, each obtained from scores derived from a video sequence. The “distance” between two sets of video-derived scores is thus learned via a discriminative classification framework. The comparison is therefore set-to-set or distribution-to-distribution in our context, rather than sample-to-sample (or score-to-score) as realised by simple fusion rules.

Although a similar direction has been pursued at the data (image) level, e.g. [6], to our best knowledge, a discriminative approach to fusion has not been attempted at video-based score level. Instead, simple fusion rules such as min, max or mean are commonly used.

A closely related approach has been reported by Tyagi et al. [13] who proposed to a discriminative training strategy, called the “maximum accept and reject” criteria, in order to improve a Bayesian score-level classifier whose underlying densities are estimated with a Gaussian Mixture Model (GMM). Other criterion such as maximal mutual information (MMI) has been used for language identification. Rather than pursuing the route of discriminative training of a generative classifier, our proposal here is to use a discriminative classifier (hence optimising a discriminative criterion) directly. Furthermore, the context is also different in that we have opted to apply our technique on a sequence of video-derived scores with arbitrary length rather than having the score dimension fixed a priori.

As for the choice of discriminative classifiers, logistic regression turns out to be very popular. For instance, it is used for fusion in combining several unimodal scores as well as other non-conventional sources of information such as biometric sample quality, cohort information, and user-dependent score characteristics [9].

Our experiments suggest that by using higher order moments, the performance can be further improved by as much as 50% compared to these simple fusion rules.

If there are two modalities, the temporal information fusion first takes place for each modality and the resultant scores are then combined at the multimodal level using the product rule (see Figure 1).

Our main contribution is the proposed novel discriminative fusion strategy that uses a vector of descriptors of the score distributions derived from the video. Although we have explored a theoretical framework using both generative and discriminative methods (to be discussed in see Section 2), the latter method is simple to implement and can be practically deployed efficiently. We implemented this method and tested it on both (talking) face and speech biometric modalities (Section 3). Experimental
evidence obtained from the BANCA (talking) face and speech video database suggests that our proposal can improve the face expert by anywhere between 7% and 30%, and the speech expert by 6%. Furthermore, the fusion of the two modality-dependent experts, after applying the temporal fusion, results in a relative improvement of 35%-50% compared to the conventional fusion (using only simple statistics to combine video-based scores), i.e., from about 2% of equal error rate to 1%.

2 A video-based multi-sample score-level fusion framework

Let $Y \in \{y_1, \ldots, y_N\}$ denote the set of matching scores $y_i \in \mathbb{R}$ derived from a video query consisting of $N$ frames of processed and valid biometric samples. For instance, for the speech modality that we will use, $N$ means the total number of Mel-scale Frequency Cepstral Coefficient (MFCC) features containing voiced speech (with the silence segments removed). For the face modality, $N$ denotes the total number of images for which our face detector can confidently find a face and the face matching algorithm can produce a score.

The most conventional strategy to obtaining a single score $y_{com}$ from the score set $Y$ is to use a simple fixed fusion rule. For the speech expert whose output is a log-likelihood ratio of two hypotheses – one hypothesis supporting that the claimant utterance comes from the “target” speaker or enrolled client versus the alternative. The speech expert that we use is a modified state-of-the-art classifier based on Gaussian Mixture Model with Maximum a posteriori adaptation (MAP-GMM) [10]. It combines the $N$ scores by using the mean rule: $y_{com} = \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$.

We employed two parts-based face experts (systems) which are very different in architecture. By parts-based we understand that a face image is represented by a set of fixed-size windows of much smaller size than the original image. The first expert uses a subset of coefficients of a Discrete Cosine Transform [3], known as DCTmod2, to represent the texture information of each sub-window. The sequence of DCTmod2 features so-derived is then classified using the MAP-GMM approach similar to the speech expert. Subsequently, the $N$ scores (from a video) are combined using the mean rule.

The second face expert represents each sub-window using non-uniform Local Binary Pattern (LBP) followed by a Fisher Linear Discriminant (FLD) projection [5]. During query, a template feature in the LBP-FLD space is compared with that of a query feature using normalised correlation. The matching scores for the respective sub-windows are then averaged to produce $y_i$ for each frame $i$ in the video. It was empirically found that the maximum rule works best to combine the scores from each video: $y_{com} = \max_{y \in Y} y$.

In the sequel, we will study a discriminative approach to combined video-derived scores. The approach departs significantly from the simple fixed rules in two ways. First, the video-based scores are now treated as a set which follows a certain distribution. Second, the distributions of client and impostor video-based scores are distinctive and repeatable for different legitimate users (also known as “target” users or clients in the speaker recognition literature or “gallery subjects” in the face recognition literature). Some training data has to be necessarily made available in order to characterise the score distributions.

Let the estimated score density of a query video, $y \in \mathcal{Y}$, be $\hat{p}(y)$. We further distinguish two types of score sets, namely, match and non-match ones, i.e., $\mathcal{Y}_k$ for $k \in \{1,0\}$. Our preliminary investigation found that the estimated densities under either the match or the non-match comparisons are very different. In particular, $\hat{p}(y)$ has a larger variance if the query video is a match rather than a non-match. We also verified that this property is consistent for both the face and the speech modalities (to be discussed in Section 3).

We shall consider a generative and a discriminative approach to classifying a query video score set, $\hat{p}(y|\mathcal{Y})$. The generative approach we considered consists of computing the following distance:

$$y_{com} = dist(\hat{p}(y|\mathcal{Y}), \hat{p}(y|\mathcal{Y}_1) - dist(\hat{p}(y|\mathcal{Y}), \hat{p}(y|\mathcal{Y}_0))$$

where $dist$ is a distance metric between two distributions. The most common choice of this metric, without making any assumption about the form of the distribution is the relative entropy (or Kullback Leibler divergence), Bhattacharyya distance, Chi-square or histogram intersection [4].

There are several difficulties when using the above generative approach. First, one has to estimate the shape of the distributions from the query data as well as the training data. This requires choosing the right form of distribution or else resorting to using a non-parametric approach such as the kernel density approach (Parzen window). Second, one needs to choose a distance metric...
between two distributions. Finally, (1) is only one possible way of comparing the merits of two distance metrics, each supporting its own hypothesis that the query video is a match or a non-match attempt.

Due to the generative nature of the above technique, many intermediate approximation steps are required, preventing us from directly minimising the classification error. For this reason, we explored a discriminative solution which aims precisely at minimising this error criterion. The idea consists of approximating the distribution of a video score \( \hat{p}(y) \) for \( y \in Y \) using simple non-parametric statistics such as mean, standard deviation, skewness, the 25-th, 50-th (median) and 75-th percentiles, as well as minimum and maximum of the scores, and the number of samples (\( N \)):

\[
\theta(Y) = [\mu, \sigma, \gamma, Q_1, Q_2, Q_3, \min(y), \max(y), N].
\]  

(2)

In essence, the above parameters summarise the entire video score set in a very coarse way. While in principal one can use many more points at different percentiles, the number of dimensions in \( \theta \) can be high, making the problem unnecessarily difficult, i.e., the curse of dimensionality.

Having derived the parameters \( \theta(Y) \), the next step consists of training a classifier using the parameters as input, estimating the posterior probability of being a genuine user, \( P(\omega_1 | \theta(Y)) \).

We used logistic regression for this purpose:

\[
P(\omega_1 | \theta(Y)) = \frac{1}{1 + \exp(-g(\theta))}
\]

(3)

where

\[
g(\theta) = \sum_r \theta_r w_r + w_0
\]

is a linear combination of the elements, \( \theta_r \), of the vector \( \theta \), defined by weights \( w_r \).

In order to train the logistic regression, we generated a training set consisting of a set of \( \theta \) each of which obtained from match score sets as positive samples. The negative samples were obtained similarly from non-match score sets, in accordance with the established experimental protocol (the BANCA pooled or “P” protocol).

Because the classifier is linear in the \( \theta \) space, the complexity of the classifier is directly related to the number of dimensions in this space. This implies that one way to increase complexity of the logistic regression is to increase the dimension of \( \theta \). This can be done, for instance, by increasing the number of percentiles describing a video score distribution. A pre-test shows that the set of parameters used in (2) is adequate, and adding more parameters with additional percentile samples does not show any significant improvement nor degradation in performance. For this reason, in the experiments to be reported in Section 3, only those parameters are used.

The discussion so far has been limited to a single expert system. In order to combine the scores of two (or more) experts, we shall introduce \( \omega \) protocol.

In order to make the accept/reject decision at the multimodal level, one simply compares \( y_{\text{final}} \) with a decision threshold. However, since our methodology also works for unimodal biometric systems, the accept/reject decision can be made by comparing \( P(\omega_1 | \theta(Y^m)) \) with a decision threshold, for each \( m \) modality independently.

3 Experiments

3.1 Expert Systems

In principle, any classifier that processes a video frame-by-frame can be used in our framework. However, in our context, we are interested in applying face and speaker verification classifiers on mobile phones where memory and computation speed are significantly reduced compared to, say, a desktop PC. For this reason, down-scaled versions of standard classifiers are used and tested here. These classifiers were implemented on Nokia N900 mobile phone [12].

The face and speaker verification baseline systems (experts) are Bayesian classifiers whose class-conditional densities are approximated using Gaussian Mixture Models (GMMs) with the Maximum a posteriori adaptation [10]. This is a well established state-of-the-art classifier for speaker verification, but since then, has also been successfully used for the face verification problem. The GMM-based face expert system that we use is reported in [3], with the source codes available at http://torch3vision.idiap.ch.

Another face expert that we used is thoroughly discussed in [5]. This expert processes each frame in a video where a face can be detected. It is worth noting that in the last Multiple Biometric Grand Challenge (MBGC) evaluation organised by NIST, a slightly more advanced version of this classifier was ranked second in the controlled evaluation setting and third in the uncontrolled setting.

The speech expert we used here differs slightly from the standard one [10] in that the speech variability across sessions is removed by factor analysis [7]. This technique is applied to all training and test data prior to building a (client-specific) GMM-MAP adapted model.

3.2 Results

We conducted our experiments in two stages: (i) unimodal experiments and (ii) multimodal experiments. In the first set of experiments, we assessed the performance of video-based frame level fusion and compared it with the standard frame-level fusion techniques. For the speech modality the so-called standard fusion is the sum rule [10]. We also confirm that this is the best strategy among all the known fixed rules. For the face modality, the max rule turns out to be the best strategy [5].

The objective of the second set of experiments is to assess the extent of improvement possible when the video-based score-level fusion is applied to the underlying expert outputs. As a control, the baseline fusion system is one whose underlying expert outputs are scores obtained on the standard fusion strategy. Therefore, the multimodal fusion module is the same (i.e., summing the expert outputs) for both systems.

The results of these two sets of experiments are shown in Table 1 in terms of Equal Error Rate (EER). This is the point at which the probability of a false accept is equal to the probability of a false reject. The third column of this table reports the
Table 1. Comparison of performance before and after score-level temporal fusion

<table>
<thead>
<tr>
<th>System</th>
<th>Before</th>
<th>After</th>
<th>Relative change (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$: IDIAP-gmm300</td>
<td>20.26</td>
<td>18.61</td>
<td>-1.64</td>
</tr>
<tr>
<td>$F_2$: IDIAP-gmm512</td>
<td>19.57</td>
<td>18.20</td>
<td>-1.37</td>
</tr>
<tr>
<td>$F_3$: UNIS-LBP</td>
<td>20.54</td>
<td>14.16</td>
<td>-6.38</td>
</tr>
<tr>
<td>$S$: UPAV-gmm</td>
<td>2.46</td>
<td>2.31</td>
<td>-0.15</td>
</tr>
<tr>
<td>$F_1 + S$ (fusion)</td>
<td>2.52</td>
<td>1.21</td>
<td>-51.11</td>
</tr>
<tr>
<td>$F_2 + S$ (fusion)</td>
<td>1.91</td>
<td>1.20</td>
<td>-37.13</td>
</tr>
<tr>
<td>$F_3 + S$ (fusion)</td>
<td>2.11</td>
<td>1.36</td>
<td>-35.51</td>
</tr>
<tr>
<td>$F_1 + F_3 + S$ (fusion)</td>
<td>2.11</td>
<td>0.98</td>
<td>-53.29</td>
</tr>
</tbody>
</table>

Note: $F_i$ denotes one of the face systems, $S$ is UPAV-gmm speaker system, and $F_i + S$ denotes a fusion system.

The relative change of EER, which is defined as

\[
\text{rel. change of EER} = \frac{\text{EER}_{\text{after}} - \text{EER}_{\text{before}}}{\text{EER}_{\text{before}}}
\]

where EER$_{\text{before}}$ is the EER before applying the proposed video-based score-level fusion (i.e., the standard fixed rule strategy) and EER$_{\text{after}}$ is the EER after applying the proposed technique. Therefore, negative change of EER implies improvement. As can be observed, although the proposed method improves only marginally the unimodal systems, the benefit is greater at the fusion level. Although EER is examined here for brevity, improvement is also observed for other operating points.

4 Conclusions

Video-based biometric systems have several advantages over their static image-based counterparts: improved robustness to spoof attack, improved accuracy via variance reduction, and the possibility of reconstructing data of higher resolution/quality. This paper explores a score-level fusion framework in which an existing biometric system can produce a matching score for each valid frame. We proposed score-level fusion strategy that relies on a set of distribution descriptors. The experimental results confirm our conjecture that a proper exploitation of the abundant scores made available by video-based biometric data can outperform the standard score-level fusion strategy. When applied to a multimodal system, as much as 50% relative gain in performance was observed (i.e., halving the EER). This result is an evidence of the merit of our proposal.

A possible extension of this work is to examine the correlation of scores between the two modalities. However, at present, this is not yet possible since the face and speech scores are not necessarily aligned. Hence, before correlation can be exploited, the alignment problem must be solved. This constitutes a possible future research direction. Another direction is to explore other discriminative classifiers such as the total error minimization approach [11].

Acknowledgements

This work was partially supported by the Mobile Biometry (MOBIO) project (IST-214324) and the Biometrics Evaluation and Testing (BEAT) project (grant no. 284989). The authors thank Dr Sebastian Marcel, Dr Chris McCool, and Dr Chi-Ho Chan for the face experts and Dr Driss Matrouf for the speech expert.

References