Global Features for Rapid Identity Verification with Dynamic Biometric Data

Morris, A.C.\textsuperscript{1}, Koreman, J.\textsuperscript{2}, Ly-Van, B.\textsuperscript{3}, Sellahewa, H.\textsuperscript{4}, Jassim, S.\textsuperscript{4}, Llarena Gómez, R.\textsuperscript{5}

\textsuperscript{1}SpinVox, Marlow, UK, \textsuperscript{2}Norwegian University of Science and Technology, Trondheim, Norway, \textsuperscript{3}Umanis, Levallois-Perret, France, \textsuperscript{4}Buckingham University, UK, \textsuperscript{5}ATOS Origin, Madrid, Spain

andrew.morris@spinvox.com, jacques.koreman@hf.ntnu.no, bao.ly_van@int-evry.fr, (harin.sellahewa,sabah.jassim)@buckingham.ac.uk, rafael.llarena@atosorigin.com

Abstract

Some of the biometrics used in identity verification, such as face, iris or fingerprint, have small fixed size pattern vectors which can easily be stored for rapid scoring. Important dynamic biometrics such as voice and signature, however, have much larger and variable sized feature vectors which require far greater verification computation. Such dynamic biometrics cannot therefore normally be used where verification must be performed rapidly on a device with little memory and very low speed. In this paper we compare the verification speed and accuracy obtained using different small fixed size global feature representations with that obtained, for voice and signature, using state of the art techniques with the full dynamic feature matrix. Using the best of these techniques we were able to reduce the multimodal processing time on a PDA SIM card, combining voice, signature and face biometrics, from one hour to ten seconds, while retaining a reduced but useful level of verification accuracy.

Index Terms: identity verification, dynamic biometrics, real time, global features, PDA, SIM

1. Introduction

In \cite{5,11} we described the implementation of a state of the art identity verification system running on a Qtek2020 PDA which uses dynamic voice and signature biometrics combined with static face features. While feature extraction took about 3 seconds on the main processor, the multimodal verification process itself took only a fraction of a second.

However, since our aim is to build a secure system that does not require any dedicated hardware, it is necessary to store biometric models and perform verification on the PDA’s SIM card, which is not accessible to manipulation and therefore allows secure storage and processing of private biometric data. With the same system as that running on the PDA verification presently takes about one hour, while a delay of more than a few seconds is not acceptable for any practical verification system. In this paper we describe how the necessary speed up was achieved while retaining a reasonable level of verification accuracy.

The root cause of the excessive processing time required for dynamic, behavioural biometric features, compared to biometrics such as iris scans or fingerprint data (as well as the face data used by us), is that state of the art authentication with voice and signature biometrics uses feature vectors sampled at 100 Hz for a variable duration of around 3 seconds (to a fixed 5 digit prompt), with each vector consisting of around 20–40 coefficients (see Section 3.1 below). The large degree of data variability also requires a complex model to represent the data (a Gaussian mixture model (GMM) \cite{2} in our case), which increases the computation required for verification scoring.

Rather than attempting to reduce the optimal sampling rate or vector size, which could not have led to anywhere approaching the required order 1000 reduction in verification time, the approach we took to this problem was to look at ways in which the patterns present in the high resolution and variable duration pattern matrix could be captured by a greatly reduced and fixed number of global features. Such a representation could also be envisaged for storage, in less secure applications, as a 2D bar code.

Behavioural biometric features are always subject to some degree of variation. Therefore, while the optimum global features used to capture such patterns would depend on the type of pattern concerned, any such pattern can be always be captured, to varying degrees of precision, by its statistical distribution, as represented by its mean, variance and higher order moments. It may be noted that for voice, the use of global features is reminiscent of the early use of long-term average spectra, LTAS \cite{8}.

In Section 2 we briefly describe the biometric data and tests used in our identity verification experiments. In Section 3 we introduce the baseline system for feature processing and multimodal verification, which we have presented elsewhere. In Section 4 we introduce the range of global features which are later tested for verification accuracy. Section 5 presents the test results which were used to select our global features. Section 6 discusses the resulting speed up and some other key implementation issues. This is followed by a conclusion.

2. Test database and protocol

To enable testing with data as close as possible to that for the PDA in real use, a database (which we refer to as the PDA database) of audio, face and signature data from 60 subjects was recorded on the PDA itself. Subjects were divided into three groups: one for Universal Background Model (UBM) training (24 subjects), and two other groups, g1 and g2 (18 subjects each). For any given FA/FR (false acceptance to false rejection) cost ratio, thresholds can then be optimised on g1 and evaluated on g2, and vice versa. In testing, data chosen from one of the speakers of the same gender and age group as the client is used to represent so-called casual impostors. A summary of this database and test protocol is given in \cite{9}.

After extensive testing it was decided that for the working system voice data would use a fixed short and psychologically neutral 5 digit prompt; that face data would use a static representation obtained from the average of feature vector over the first 10 video frames (while the mouth is usually still closed), and that for enrolment the prompt should be read and a signature written 8 times under varying conditions of background noise, acoustics, lighting and touch screen stability.
3. Baseline verification system

3.1. Feature data

As the user reads a fixed prompt, audio data is sampled at 22 kHz and the facial image at 20 frames per second. The user’s signature is then read at 100 x-y points per second (commonly used angle and pressure are not available on a PDA).

Voice: 19 Mel-frequency cepstral coefficients (MFCCs), plus first time difference features, were computed every 10ms from a 20ms window, using a pre-emphasis factor of 0.97, a Hamming window and 20 Mel-scaled filters (ignoring c0). Features were obtained using HTK [13].

Face: After histogram equalisation to normalise for variation in lighting conditions, 120 LL4 Haar wavelet features [12] were computed for the first 10 video frames.

Signature: To the x-y data were added derived features of velocity, acceleration, curvature and others [5]), to give a total of 19 features per point.

The above details resulted from extensive optimisation and give state of the art performance for each modality.

3.2. Multimodal verification

The baseline verification system running on the PDA (i.e., without use of the SIM card) uses a separate state-of-the-art GMM to model the feature probability distribution for each of the three biometric modalities [10].

3.3. Match scoring

With this system a universal background model (UBM) is first trained on features from a number of other clients. The UBM is then used as an impostor model, and a client model is trained by MAP adaptation [7] of just the Gaussian means of the UBM to the limited client data available. Diagonal covariance GMMs (1) were trained and tested using the Torch machine learning API [1]. If 

\[ p(x_i | C) \Rightarrow \sum_{i=1}^{N} \alpha_i N(x_i, \mu_i, \Sigma_i) \]  

\[ \log p(X_i | C) = \frac{1}{T} \sum_{i=1}^{T} \log p(x_i | C) \]

\( \mathcal{N}(\cdot) \) in (1) is the multivariate Normal (Gaussian) distribution. \( X_i \) in (2) is the full dynamic feature matrix, whose log likelihood is estimated as the average frame log likelihood. GMM training used k-means clustering followed by EM iteration. Optimisation resulted in GMMs for voice and signature using 100 Gaussians, and for face using just 4 Gaussians. The match score for each modality is then given by the logarithm of the likelihood ratio (3) (which is proportional to the ratio of client to impostor posterior probabilities).

\[ s_i = \log p(X_i | C) - \log p(X_i | I) \]  

3.4. Scores combination

The separate scores (3) from each modality are combined by modeling both the joint client scores likelihood and the joint impostor scores likelihood by separate GMMs, and then obtaining the combined score as the client posterior probability, using Bayes’ rule (4)

\[ P(C|s_1, s_2, s_3) = \frac{P(C)p(s_1, s_2, s_3 | C)}{P(C)p(s_1, s_2, s_3 | C) + P(I)p(s_1, s_2, s_3 | I)} \]  

Here the client prior probability \( P(C) \) was set to 0.5, and \( P(I) = 1-P(C) \). The claimed identity is accepted if this combined score is above some preset threshold.

3.5. Baseline verification accuracy

The baseline system described ran successfully on the PDA main processor, with the verification computation taking only a fraction of a second (a demonstration video of the system is available on the SecurePhone web pages www.securephone.info). Baseline test results are given in Table 1.

<table>
<thead>
<tr>
<th>Modality</th>
<th>%EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voice</td>
<td>7.21</td>
</tr>
<tr>
<td>Face</td>
<td>28.40</td>
</tr>
<tr>
<td>Signature</td>
<td>8.01</td>
</tr>
<tr>
<td>Fusion</td>
<td>2.39</td>
</tr>
</tbody>
</table>

Table 1. %EER for voice (using a 5 digit prompt), face and signature, and the combination of these three modalities

However, while this system conformed to the memory limitations that would be required to run on the SIM processor, when implemented on the SIM, verification took over an hour to run.

4. Global feature representations tested

While face verification uses only the first video frames, the state of the art speaker and signature verification system presented in the previous section trains a client GMM on the full set of biometric feature frames. One multi-dimensional Gaussian is associated with each vector cluster and it is the distribution of these clusters which distinguishes each person.

4.1. Statistical moments

To reduce the number of dynamic biometric feature vectors (frames), we initially simply replaced the sequence of T feature vectors from each utterance or signature (a cloud of points) by a single point at the average vector position. Although this single point tells us little about the original cluster distribution which normally serves to characterise each speaker (for speech it is comparable to LTAS; for the signature it should tell us nothing at all), some further information about the original feature distribution can be provided by the corresponding vector of variances (so representing the cloud by an ellipsoid), and by further higher order statistics. Furthermore, if we divide the frame sequence into Nv equal parts and compute separate mean and variance vectors for each part, we can get closer to the original unglobalised feature vector sequence.

4.2. Tailor made features

Knowledge of the kinds of patterns which a particular biometric might exhibit could allow for tailor made global features which are optimal for this modality. For signature mode one such tailor made global representation was tested, as well as the global mean and standard deviation (GMSD) vectors. The representation tested, which we will refer to as GSIG (global signature) features, consisted of 41 features...
which included average position, velocity and acceleration; their variances; their x-y correlations; the number of sign changes and runs with the same sign in velocity (number and length of strokes), as well as means and variances computed from the normalised parameters [6].

5. Test results

Test results for the system presented in the previous section were obtained on the PDA (which has floating point emulation), not on the SIM (which does not). However, integerisation should not have a major effect on accuracy.

<table>
<thead>
<tr>
<th>Np</th>
<th>mean</th>
<th>mean + st. dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28.20</td>
<td>30.08</td>
</tr>
<tr>
<td>2</td>
<td>22.13</td>
<td>21.09</td>
</tr>
<tr>
<td>3</td>
<td>21.64</td>
<td>19.31</td>
</tr>
<tr>
<td>4</td>
<td>21.16</td>
<td>20.31</td>
</tr>
<tr>
<td>8</td>
<td>37.26</td>
<td>29.28</td>
</tr>
<tr>
<td>1</td>
<td>28.29</td>
<td>27.15</td>
</tr>
<tr>
<td>2</td>
<td>28.29</td>
<td>27.58</td>
</tr>
<tr>
<td>3</td>
<td>38.30</td>
<td>27.47</td>
</tr>
<tr>
<td>4</td>
<td>37.24</td>
<td>25.73</td>
</tr>
<tr>
<td>8</td>
<td>32.26</td>
<td>31.78</td>
</tr>
</tbody>
</table>

Table 2. %EER for voice, signature and face alone, each using global mean or global GMSD features, and after scores fusion. Voice and signature are constrained to use the same number of parts (Np) and of Gaussians (Ng).

The results above clearly show the advantage of combining the scores from three modalities. Even if the global features for individual modalities have EERs which are much higher than dynamic features (cf. voice and signature results in Table 2 with those in Table 1), the best result for global features (7.30% EER) is fairly good — although still below that of the reference system presented in Table 1 (2.39% EER).

For practical purposes, the system we implemented on the SIM card uses truly global features for the voice and signature modalities (Np=1). This reduces the verification time, making the system very user-friendly, while retaining an EER of 10.05%, which is only four times that of our baseline system.

City-block distance (CBD, sum of absolute vector differences) was tested as an alternative to GMM distance (Table 3). Both can be obtained using only simple arithmetic operations (see Eq. 5, Section 6.4), but CBD computation is quicker by a factor of 2 and can sometimes have advantageous behaviour. Min and Avg CBD are the minimum and average CBD distance from a vector to each of several template vectors. Although CBD leads to a much lower performance in each individual modality, the fusion result is surprisingly close to that of GMM (and indeed better than GMM when GSIG signature features are used with the global means only).

The use of GSIG features for the signature, presented in the lower part of Table 3, does not lead to an improvement and its pre-processing is computationally more intensive.

<table>
<thead>
<tr>
<th></th>
<th>GMM</th>
<th>Min CBD</th>
<th>Avg CBD</th>
<th>GMM</th>
<th>Min CBD</th>
<th>Avg CBD</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMSD global voice and face features</td>
<td>30.36</td>
<td>68.90</td>
<td>-</td>
<td>24.41</td>
<td>73.34</td>
<td>70.76</td>
</tr>
<tr>
<td>- combined with GSIG global signature features</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>10.05</td>
<td>12.61</td>
<td>15.56</td>
</tr>
<tr>
<td>Sign.</td>
<td>7.30</td>
<td>70.49</td>
<td>-</td>
<td>21.27</td>
<td>76.03</td>
<td>70.72</td>
</tr>
<tr>
<td>Fused</td>
<td>14.83</td>
<td>18.22</td>
<td>-</td>
<td>7.30</td>
<td>17.02</td>
<td>16.66</td>
</tr>
</tbody>
</table>

Table 3. Comparison of the use of GMM and CBD, with global features as mean vector or mean plus standard deviation vector (Nv=1, Ng=4). Face always uses mean only. %EER scores are shown for each modality separately and after fusion with GMSD or GSIG signature features.

6. Discussion

The aim of this paper was to present a secure and practical identity verification system which can run on a standard PDA. The face, voice and signature modalities which the system uses are user-friendly and natural, given that the application runs on a PDA. Unlike the face, however, the voice and online signature consist of highly variable behavioural data which require large GMMs to model them and a strong processor to perform verification in real time. This clashes with the requirement that the biometric models must be stored and verified on the SIM card of the PDA for security reasons.

6.1. GMSD features provide required speed up

The problem was solved, at least as far as the system’s speed is concerned, by using the simple mean and variance over time as global features for the voice and signature. These reduce the feature representation to a single frame instead of one frame for every 10 ms. These features also greatly reduce feature variation (hence model size) while retaining a reasonable power of discrimination. While performance dropped from under 2.5% EER to around 10% for the implemented system, as global features smooth out fine detail they are likely to be considerably more robust to noise.

Accuracy could of course be improved by using dedicated iris or fingerprint sensors, but these solutions would be more costly, while a server solution would raise privacy issues because the biometric data could be intercepted. These solutions would not fulfill the basic requirements of our work: high user acceptance of the modalities used for identity verification, lack of special hardware requirements and very high privacy standards.

6.2. GMSD features provide model size reduction

As well as reducing the number of feature frames which required processing for GMM evaluation by a factor of about
300, the fact that globalised features have far less variation than local features resulted in a reduction in the optimum number of Gaussians for both voice and signature biometrics, from 100 to just 4. This provided a further factor of 25 reduction in model size. After global mean and variance calculation, this resulted in an overall speed up in verification computation by a factor of at least 3000. Feature globalisation also reduced the time required to transmit features from the PDA to the SIM from around 4.5 minutes to just a few milliseconds.

6.3. Online processing

Feature processing is not privacy-sensitive and can thus be performed on the PDA itself. While the PDA processor is fast, feature processing requires thousands of times more processing than verification scoring. To maximise the speed of the authentication response, online versions of silence detection and CMS (Cepstral Mean Subtraction used with MFCC generation) were both implemented for speech pre-processing, with silence detection using a running estimate of the additive noise level and CMS using a running estimate of the convolutive noise level. Online CMS showed improved performance over offline CMS with PDA data. Processing time was therefore further reduced by starting the verification process before audio-visual acquisition is completed.

6.4. Integerisation

Most SIM cards, including our selected SIM card, have only integer arithmetic operations. All model parameters and feature data passed to the SIM were integerised by linear projection onto a fraction of the observed range for each feature value. Evaluation of the logarithm of the GMM function (2) and (3) therefore required use of the well known “maximum approximation”, which requires no log operation.

\[
\log \sum_{i=1}^{N} \alpha_i N(x_i, \mu_i, \Sigma_i) \cong \max \left[ \log \left( \alpha_i N(x_i, \mu_i, \Sigma_i) \right) \right] \quad (5)
\]

6.5. PDA-SIM communication

Two communication models were used to pass the voice, face and signature features on to the SIM card for subsequent verification: one for a card working in the PDA internal reader and another to be used with an external reader (to enable solutions which are not GSM-based). To communicate with an external reader, simple APDU commands can be used. To send APDUs to the SecurePhone module inside the internal card, however, it is necessary to select that module, deselecting the card’s GSM module and losing GSM connectivity. Since the SecurePhone aim is to perform verification and document exchange during a phone call, this must be avoided. To this end, we implemented a communication protocol based on the SIM Application Toolkit (SAT) [3]. To adapt the simple APDU protocol implemented to the SAT standard, complete APDU packets were inserted in the data field of an SMS-PP data download envelope. In this way, feature data can be sent to the internal SIM card without losing GSM connectivity.

7. Conclusion

Using the simple mean and standard deviation vectors over time as global features we were able to reduce processing by a factor of 3750 (300 dynamic frames to 1 mean and 1 variance vectors, and 100 Gaussians to 4), to enable the SecurePhone verification application to run in real-time on a standard SIM card. Although performance was reduced from 2.4% to around 10% EER, the SecurePhone serves as a proof of concept for the usability of PDA based multi-modal identity verification.

Acknowledgements

This work was closely supported by all of the partners in this EC SecurePhone project IST-2002-506883.

References