IDIAP RESEARCH REPORT

INTER-SESSION VARIABILITY MODELLING AND JOINT FACTOR ANALYSIS FOR FACE AUTHENTICATION

Roy Wallace Mitchell McLaren Chris McCool
Sébastien Marcel

Idiap-RR-28-2011

AUGUST 2011

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Inter-session Variability Modelling and Joint Factor Analysis for Face Authentication

Roy Wallace
Idiap Research Institute, Martigny, Switzerland
roy.wallace@idiap.ch

Mitchell McLaren
Radboud University Nijmegen, The Netherlands
m.mclaren@let.ru.nl

Christopher McCool, Sébastien Marcel
Idiap Research Institute, Martigny, Switzerland
christopher.mccool@idiap.ch, sebastien.marcel@idiap.ch *†‡

Abstract

This paper applies inter-session variability modelling and joint factor analysis to face authentication using Gaussian mixture models. These techniques, originally developed for speaker authentication, aim to explicitly model and remove detrimental within-client (inter-session) variation from client models. We apply the techniques to face authentication on the publicly-available BANCA, SCface and MOBIO databases. We propose a face authentication protocol for the challenging SCface database, and provide the first results on the MOBIO still face protocol. The techniques provide relative reductions in error rate of up to 44%, using only limited training data. On the BANCA database, our results represent a 31% reduction in error rate when benchmarked against previous work.

1. Introduction

Many challenges in face authentication can be attributed to the problem of session variability, that is, changes in environment, illumination, pose, expression, or image acquisition, which cause mismatch between images of the same client (person). While face authentication has evolved considerably in the past 15 years, modern approaches still suffer from increased errors in the presence of substantial session variability [19, 1].

One approach of particular interest uses a parts-based topology, whereby the distribution of features extracted from images of a client's face is described by a Gaussian mixture model (GMM) [20]. In [7, 6], this approach was found to offer the best trade-off in terms of complexity, robustness and discrimination. Interestingly, a similar GMM-based approach [17] forms the basis of state-of-the-art speaker authentication, in which the issue of session variability has already received considerable focus [11, 24, 15].

Two of the most successful techniques in improving robustness to session variability for speaker authentication are inter-session variability modelling (ISV) and the related technique of joint factor analysis (JFA), which have been shown to reduce errors by more than 30% [11, 24, 15]. ISV and JFA aim to estimate more reliable client models by explicitly modelling and removing within-client variation using a low-dimensional subspace. In speaker authentication, this detrimental variation is caused by different microphones, acoustic environments and transmission channels. JFA can be considered to be an extension of ISV as it also explicitly models between-client variation.

In this paper, we apply ISV and JFA to the face authentication task. The intuition is that the sources of within-class variation in speech have parallels in facial images. Specifically, face authentication performance is adversely affected by the effects of variation in environment, expression, pose and image acquisition, which we hypothesise can be modelled and removed using ISV and JFA. This hypothesis is supported by this work, as the ISV and JFA techniques reduce the error rate by between 11% and 44% across the challenging BANCA, SCface and MOBIO databases.

Section 2 introduces GMM-based face authentication. Then, ISV and JFA are described in Section 3. In Sections 4, 5 and 6, we present and discuss results on the BANCA, SCface and MOBIO databases.

*The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7) under grant agreements 238803 (BBfor2) and 257289 (TABULA RASA).

†Thanks to Ondřej Glembek, Lukáš Burget and Pavel Matějka from the FIT group at BRNO University of Technology for their advice and assistance in the implementation of the JFA system used in this work.

‡Portions of the research in this paper use the SCface database of facial images. Credit is hereby given to the University of Zagreb, Faculty of Electrical Engineering and Computing for providing the database of facial images.

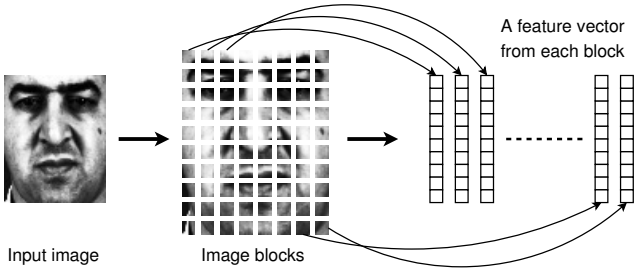


Figure 1: The concept of a parts-based topology: dividing the face into blocks and obtaining a feature vector from each block.

2. GMMs for face authentication

The GMM parts-based topology was first applied to face authentication in [20] and has since been successfully utilised by several researchers [12, 6, 13]. The method decomposes the face into a set of blocks that are considered to be separate observations of the same signal (the face). An overview of the procedure is shown in Figure 1. The main difference between this approach and the similar GMM-based approach to speaker authentication is the use of visual features extracted from parts of the face image, rather than acoustic features extracted from frames of the speech signal.

Since features are extracted from each part of the face independently, the approach is naturally robust to occlusion, local transformation and face mis-localisation, and has been found to offer the best trade-off in terms of complexity, robustness and discrimination [7, 6]. The rest of this section describes the main processing stages of the framework, including image pre-processing, feature extraction and classification.

2.1. Image pre-processing and feature extraction

Each image is rotated, cropped and registered to a 64×80 intensity image with the eyes 16 pixels from the top and separated by 33 pixels. Optionally, each cropped image is processed using Tan & Triggs normalisation [21]. From each normalised image we exhaustively sample $B \times B$ blocks of pixel values by moving the sampling window one pixel at a time. Each block is mean and variance normalised prior to extracting the D lowest-frequency 2D-DCT coefficients. The resulting feature vectors extracted from an image are finally mean and variance normalised in each dimension. Each image is thus represented by a set of K feature vectors, $\boldsymbol{O} = \{\boldsymbol{o}^1, \boldsymbol{o}^2, \ldots, \boldsymbol{o}^K\}$.

2.2. Modelling and Classification

The distribution of feature vectors for each client is modelled by a Gaussian mixture model, estimated using background model adaptation [6, 12, 17]. Background model adaptation utilises a universal background model (UBM), $\boldsymbol{m}$, as a prior for deriving client models using maximum *a posteriori* (MAP) adaptation [17]. We only adapt the means of the GMM components and use diagonal covariance matrices, as this requires fewer observations to perform adaptation [17] and has already been shown to be effective for face authentication [12, 6].

A test image, $\boldsymbol{O}_t$, can then be verified by scoring against the model of the claimed client identity ($\boldsymbol{s}_i$) and the UBM ($\boldsymbol{m}$). The two models, $\boldsymbol{s}_i$ and $\boldsymbol{m}$, each produce a log-likelihood from which a log-likelihood ratio (LLR) is calculated,

$$h\left(\boldsymbol{O}_t, \boldsymbol{s}_i\right) = \sum_{k=1}^{K} \left(\log(p(\boldsymbol{o}_t^k \mid \boldsymbol{s}_i)) - \log(p(\boldsymbol{o}_t^k \mid \boldsymbol{m}))\right), \tag{1}$$

to produce a single score. The image $\boldsymbol{O}_t$ is then classified as belonging to client i if and only if $h\left(\boldsymbol{O}_t, \boldsymbol{s}_i\right)$ is greater than a threshold, θ.

In this work, we use a fast scoring technique known as *linear scoring* [8]. This relies on using a particular representation of a GMM called a *GMM mean supervector*, which is formed by concatenating the GMM component means. Thus, with GMMs expressed as supervectors, linear scoring approximates (1) with

$$h_{\text{linear}}\left(\boldsymbol{O}_t, \boldsymbol{s}_i\right) = \bar{\boldsymbol{s}}_i^{\top} \boldsymbol{\Sigma}^{-1} \bar{\boldsymbol{F}}_t, \tag{2}$$

where $\bar{\boldsymbol{s}}_i = \boldsymbol{s}_i - \boldsymbol{m}$, $\boldsymbol{\Sigma}$ is the diagonal matrix formed by concatenating the diagonals of the UBM covariance matrices, $\bar{\boldsymbol{F}}_t = \boldsymbol{F}_t - \boldsymbol{N}_t \boldsymbol{m}$ is the supervector of UBM-centralised first order UBM statistics of the image, and $\boldsymbol{N}_t$ is the diagonal matrix formed by concatenating the zeroth order UBM statistics of the test image. Readers are referred to [8] for more details. Finally, the scores are normalised using ZT-norm (Z-norm followed by T-norm) [2].

3. Session variability modelling

Inter-session variability modelling (ISV) and joint factor analysis (JFA) are two *session variability modelling* techniques that have been applied with success to speaker authentication. This section provides a brief overview of these techniques. Readers are referred to [11, 24, 15] for more details.

Session variability modelling aims to estimate and exclude the effects of within-client variation, in order to create more reliable client models. At enrolment time, a GMM is trained for each client by adapting the means of a UBM, as described in Section 2. In terms of GMM mean supervectors, client enrolment can be expressed as

$$\boldsymbol{s}_i = \boldsymbol{m} + \boldsymbol{d}_i, \tag{3}$$

where $\boldsymbol{m}$ is the UBM mean supervector, $\boldsymbol{d}_i$ is the client-dependent offset, and $\boldsymbol{s}_i$ is the resulting model for client i. Ideally, the resulting client model should be robust to any variations within the client's enrolment images due to, for example, changes in illumination, expression or pose. However, this variation is not accounted for in (3), and it is therefore likely that this will lead to a suboptimal client model, particularly in the case of limited enrolment data.

Session variability modelling proposes to explicitly model the variation between different sessions of the same client and exclude this variation from the client models during enrolment as well as testing. In our case, we consider that each image was acquired during a different session. The particular conditions of a session are assumed to result in an offset to each of the GMM component mean vectors [24],

$$\boldsymbol{\mu}_{i,j} = \boldsymbol{s}_i + \boldsymbol{u}_{i,j}, \tag{4}$$

where $\boldsymbol{u}_{i,j}$ is the session-dependent offset for the j'th image of client i, and $\boldsymbol{\mu}_{i,j}$ is the resulting mean supervector of the GMM that best represents the image ($\boldsymbol{O}_{i,j}$). The goal of enrolment using session variability modelling is to find the true session-independent client model, $\boldsymbol{s}_i$, by jointly estimating this along with each $\boldsymbol{u}_{i,j}$. Techniques for this estimation will be discussed in the following sections.

At test time, in contrast to (1), the score for an image $\boldsymbol{O}_t$ is calculated as

$$h\left(\boldsymbol{O}_t, \boldsymbol{s}_i\right) = \sum_{k=1}^{K} \big(\log(p(\boldsymbol{o}_t^k \mid \boldsymbol{s}_i + \boldsymbol{u}_{i,t})) - \log(p(\boldsymbol{o}_t^k \mid \boldsymbol{m} + \boldsymbol{u}_{\text{UBM},t})) \big) \tag{5}$$

where $\boldsymbol{u}_{i,t}$ and $\boldsymbol{u}_{\text{UBM},t}$ are the MAP estimates of the session-dependent offsets for the image estimated using the client and UBM models respectively[1]. In practice, this LLR is approximated using linear scoring. By removing the estimated session offset from the first order statistics, (2) thus becomes

$$h_{\text{linear}}\left(\boldsymbol{O}_t, \boldsymbol{s}_i\right) = \bar{\boldsymbol{s}}_i^{\top} \boldsymbol{\Sigma}^{-1} \left(\bar{\boldsymbol{F}}_t - \boldsymbol{N}_t \boldsymbol{u}_{\text{UBM},t}\right). \tag{6}$$

3.1. Inter-session variability modelling (ISV)

The ISV technique, proposed in [24], assumes that within-client variation is contained in a linear subspace of the GMM mean supervector space. That is,

$$\boldsymbol{u}_{i,j} = \boldsymbol{U}\boldsymbol{x}_{i,j}, \tag{7}$$

where $\boldsymbol{U}$ is the low-dimensional subspace that contains within-client variation, and $\boldsymbol{x}_{i,j} \sim \mathcal{N}(0, \boldsymbol{I})$. The client-dependent offset (see (3)) is set to

$$\boldsymbol{d}_i = \boldsymbol{D}\boldsymbol{z}_i, \tag{8}$$

[1] In practice, the session offsets are assumed to be identical both from the client model and the UBM, so that $\boldsymbol{u}_{i,t}$ is approximated by $\boldsymbol{u}_{\text{UBM},t}$. This is referred to as the *LPT assumption* in [8].

where $\boldsymbol{D}$ is a diagonal matrix with elements

$$\boldsymbol{D}(q,q) = \sqrt{\frac{\boldsymbol{\Sigma}(q,q)}{\tau}} \tag{9}$$

and $\boldsymbol{z}_i \sim \mathcal{N}(0, \boldsymbol{I})$. Here, $\boldsymbol{\Sigma}(q,q)$ is the variance from the UBM and τ is the adaptation relevance factor [17]. The matrix $\boldsymbol{D}$ is set in this way to ensure that the MAP solution for $\boldsymbol{d}_i$ in (3) is equivalent to the classical MAP mean update rule of [17].

To summarise, each image is represented by a GMM mean supervector

$$\boldsymbol{\mu}_{i,j} = \boldsymbol{m} + \boldsymbol{U}\boldsymbol{x}_{i,j} + \boldsymbol{D}\boldsymbol{z}_i. \tag{10}$$

Thus, by explicitly modelling the session-dependent offsets $\boldsymbol{u}_{i,j} = \boldsymbol{U}\boldsymbol{x}_{i,j}$, the aim is to exclude these effects of session variability from the resulting client models, $\boldsymbol{s}_i = \boldsymbol{m} + \boldsymbol{D}\boldsymbol{z}_i$.

3.2. Estimation of subspaces and latent variables

To use the ISV framework we need to be able to (i) estimate the latent variables, $\boldsymbol{x}_{i,j}$ and $\boldsymbol{z}_i$, and (ii) train the subspace $\boldsymbol{U}$. In general, MAP estimation is used to solve problem (i) and maximum likelihood (ML) is used to solve problem (ii).

Our approach is based on the algorithms described in [24]. The latent variables, $\boldsymbol{x}_{i,j}$ and $\boldsymbol{z}_i$, are jointly estimated using MAP estimation. First, define the set of latent variables for client i's J enrolment images as $\bar{\boldsymbol{\lambda}}_i = \{\boldsymbol{z}_i, \boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, \ldots, \boldsymbol{x}_{i,J}\}$, then

$$\bar{\boldsymbol{\lambda}}_i = \operatorname*{argmax}_{\boldsymbol{\lambda}_i} p(\boldsymbol{\lambda}_i \mid \boldsymbol{O}_{i,1}, \boldsymbol{O}_{i,2}, \ldots, \boldsymbol{O}_{i,J}), \tag{11}$$

$$= \operatorname*{argmax}_{\boldsymbol{\lambda}_i} p(\boldsymbol{z}_i) \prod_{j=1}^{J} p(\boldsymbol{O}_{i,j} \mid \boldsymbol{x}_{i,j}, \boldsymbol{z}_i) p(\boldsymbol{x}_{i,j}). \tag{12}$$

This is solved using the Gauss-Seidel approximation method of [24]. The subspace $\boldsymbol{U}$ is trained on background data using the ML-based iterative method described in Section 5.2 of [24], which alternates between ML updates of $\boldsymbol{U}$ and MAP estimation of the latent variables, as described above.

3.3. Joint factor analysis (JFA)

The JFA modelling technique [11] can be seen as an extension of ISV. Specifically, in contrast to (8), the client-dependent offset is defined as

$$\boldsymbol{d}_i = \boldsymbol{V}\boldsymbol{y}_i + \boldsymbol{D}\boldsymbol{z}_i, \tag{13}$$

where $\boldsymbol{V}$ is a rectangular matrix of low rank, $\boldsymbol{y}_i \sim \mathcal{N}(0, \boldsymbol{I})$, and $\boldsymbol{d}_i$ is thus distributed with covariance matrix $\boldsymbol{D}^2 + \boldsymbol{V}\boldsymbol{V}^{\top}$. The assumption of this model is that most between-client variability is contained within a low-dimensional subspace $\boldsymbol{V}$, which is in fact the assumption of the well-known

eigenvoice [22] and eigenface [23] modelling techniques. One of the motivations for using JFA is to improve enrolment with limited data, by allowing a client model to be approximately represented by only the small number of factors in $\boldsymbol{y}_i$. To summarise, in contrast to (10), for JFA each image is modelled as

$$\boldsymbol{\mu}_{i,j} = \boldsymbol{m} + \boldsymbol{U}\boldsymbol{x}_{i,j} + \boldsymbol{V}\boldsymbol{y}_i + \boldsymbol{D}\boldsymbol{z}_i. \quad (14)$$

In this case, both $\boldsymbol{V}$ and $\boldsymbol{D}$ are learnt from training data, in addition to $\boldsymbol{U}$, using maximum likelihood [11]. As with ISV, JFA aims to exclude the effects of session variability, such that the resulting client models are $\boldsymbol{s}_i = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y}_i + \boldsymbol{D}\boldsymbol{z}_i$.

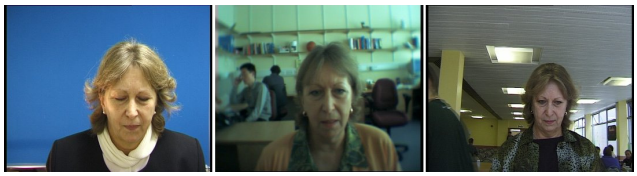
4. Experimental protocols

To assess face authentication accuracy, scores were first generated on a development set, from which a global decision threshold was found that minimised the equal error rate (EER). This threshold was then applied to a test set of completely separate clients to find the half total error rate (HTER), that is, the average of false acceptance and false rejection rates. Thus the threshold, as well as all other hyperparameters, were tuned prior to seeing the test set. This is a critical requirement if such technology is to be applied to real applications [5].

In the past, a wide variety of databases have been used for evaluation of face authentication techniques. To properly evaluate ISV and JFA, we chose to use images taken in challenging conditions causing substantial within-client variation. Furthermore, we chose to restrict ourselves to publicly-available databases with separate training, development and test sets to allow for unbiased evaluation. Unfortunately, some popular databases such as FRGC [16] and LFW [10] were thus not applicable, as they do not include separate development and test sets[2]. We therefore chose to evaluate the ISV and JFA techniques on the challenging BANCA, SCface and MOBIO databases. The BANCA and MOBIO databases already have well-defined protocols, while a suitable face authentication protocol for SCface is proposed in this work.

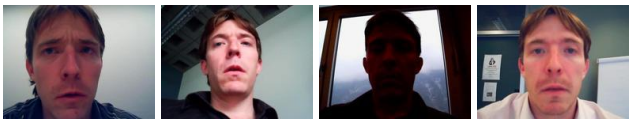
4.1. BANCA

Results are reported for the Pooled test (P) on the English subset of the BANCA database [3]. While BANCA is actually a multi-modal database of videos, we used the 5 pre-selected still images from each video and treated each image independently, as specified in the protocol. Images were captured in three different scenarios, referred to as *controlled*, *degraded* and *adverse*. As shown in Figure 2a, significant session variation exists between images of the same client.



(a) BANCA database: controlled, degraded, and adverse scenarios.

(b) SCface database: enrolment/mugshot image; test images from close, medium and far distances.



(c) MOBIO database.

Figure 2: Example images showing a wide range of within-client variation (session variability).

The *g1* and *g2* groups of clients (26 each) were used as the development and test sets respectively. For both sets, each client model was enrolled with 5 images from the controlled scenario. Then, a total of 2,730 scores (1,170 target trials, 1,560 impostor trials) were generated using images across all three scenarios. The UBM was trained from 200 images of 20 clients in the separate *world data* set. All world data was used to train $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{D}$ (300 images of 30 clients). As score normalisation is more effective when using a set of clients disjoint from those in the UBM training set, *g1* was used to normalise the scores for *g2*, and vice-versa. Images were cropped using either manually annotated eye locations, or automatic face localisation based on the detector of [18][3].

4.2. SCface

The Surveillance Cameras face database (SCface) [9] is particularly interesting from a forensics point-of-view because images were acquired using commercially available surveillance equipment, in a range of challenging but realistic conditions. As could be imagined in a real-life scenario, for authentication, these surveillance images are compared to a single high-resolution mugshot image.

[2] In the FRGC database, 153 clients occur in both the training set as well as the test set, and there is no publicly-available development set. In the LFW database, 758 image pairs in the training/development set (*View 1*) are exactly repeated in the test set (*View 2*).

[3] Implemented with Torch3vision (http://torch3vision.idiap.ch).

While [9] suggested a protocol for face *identification*, it did not include a world data set or a development data set, and no protocol was suggested for face authentication. Therefore, we propose the following face authentication protocol for SCface based on the *DayTime tests* scenario [9][4]. The database was divided into subsets based on client ID such that clients 1–43, 44–87, and 88–130 were allocated to world data, development, and test sets, respectively. Each client model was enrolled using a single mugshot image. Then, test images were taken from the 5 surveillance cameras at 3 different distances: *close*, *medium* and *far*. Each client model was tested against the 15 surveillance images from each client in the same subset. This resulted in 645 target trials and 27,090 impostor trials in the test set. Unless otherwise noted, results are reported for a *combined* protocol, in which each test image was assumed to originate from an unknown camera at an unknown distance. Two-thirds of the world data was used for UBM training, the other third was used for score normalisation, while all world data was used to train $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{D}$ (688 images of 43 clients). During pre-processing, manually annotated eye locations were used for cropping, and low resolution images were upsampled where necessary.

Example mugshot and surveillance images are provided in Figure 2b. From the figure, considerable session variation in terms of quality, pose and illumination can be observed. It is this variation that we attempt to explicitly model and remove in the following experiments.

4.3. MOBIO

The MOBIO database is a large and challenging biometric database. It contains videos of 150 participants captured in challenging real-world conditions on a mobile phone camera over a one and a half year period [14]. Figure 2c shows example images, demonstrating session variability due to variation in pose and illumination. For this work, one image was extracted from each of the videos and was manually annotated with eye locations. Using manual face localisation allows us to evaluate face authentication accuracy separately from the choice of face detection algorithm. These images and annotations will be made available to facilitate future benchmarking[5].

The MOBIO protocol is supplied with the database and defines three non-overlapping partitions: training, development and testing. The development and testing partitions are defined in a gender-dependent manner, such that clients' models are only tested against images from clients of the same gender. We chose to use the training data in a gender-independent manner to be consistent with the other databases, though future work could investigate gender-dependent training. For male clients, the test set contains 3,990 target trials and 147,630 impostor trials. For female clients, there are 2,100 target trials and 39,900 impostor trials. All of the training data was used to train $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{D}$ (9,579 images of 50 clients). A subset of 1,224 images of 34 clients (36 images each) was used for UBM training, while the other 16 clients were used for score normalisation.

[4] http://scface.org/

[5] http://www.idiap.ch/dataset/mobio

5. Results

In this section, results are reported for each database independently. A GMM parts-based system, as described in Section 2, without ISV or JFA is used as the baseline system for comparison. Hyper-parameters were tuned on the development set for each database, including the block size used during feature extraction, and the dimensionality of subspaces $\boldsymbol{U}$ and $\boldsymbol{V}$. UBMs were trained with 512-components[6], and a relevance factor of $\tau = 4$ was used for client model adaptation. For experiments on SCface only, cropped images were not pre-processed using Tan & Triggs normalisation [21], as it did not improve performance in that case. ISV and JFA were implemented based on the JFA cookbook[7]. Subspaces $\boldsymbol{V}$, $\boldsymbol{U}$ and $\boldsymbol{D}$ were trained using 10 EM iterations, in that order for JFA. For ISV, only $\boldsymbol{U}$ was trained. Latent variables were estimated in the order $\boldsymbol{y}_i$ (JFA only), $\boldsymbol{x}_{i,j}$, then $\boldsymbol{z}_i$, using one Gauss-Seidel iteration [24]. Manual face localisation was utilised unless otherwise noted.

For evaluating the statistical significance of improvements in HTER, we used the methodology proposed by equation (15) and Figure 2 of [4], with a one-tailed test.

5.1. Feature extraction and score normalisation

Firstly, the size of the blocks used during feature extraction was tuned on development data. The number of DCT coefficients for a given block size, D, was initially tuned on BANCA. As shown in Table 1, the optimal block sizes were 12×12 and 20×20 pixels for the BANCA and SCface databases, respectively. For MOBIO, across male and female clients, a block size of 12×12 was chosen.

Table 2 illustrates that ZT-norm score normalisation was very effective for BANCA and SCface, with relative reductions in test set HTER of 45% and 35% respectively. For MOBIO, ZT-norm had little effect.

This system, with tuned block size and ZT-norm score normalisation but without ISV or JFA, is referred to as the baseline system for the following session variability modelling experiments.

5.2. Session variability modelling on BANCA

Table 3 compares the ISV and JFA session variability modelling techniques to the baseline approach on BANCA.

[6] Using the Torch3vision library (http://torch3vision.idiap.ch/)

[7] Available at: http://speech.fit.vutbr.cz/en/software/joint-factor-analysis-matlab-demo

B	D	BANCA		SCface		MOBIO (male)		MOBIO (female)	
		Dev	Test	Dev	Test	Dev	Test	Dev	Test
8	28	9.3%	8.2%	23.8%	25.7%	10.8%	11.1%	**10.6%**	**19.5%**
12	45	**7.8%**	**6.1%**	20.2%	20.6%	**9.2%**	**10.5%**	10.7%	20.4%
16	66	**7.8%**	6.5%	18.5%	17.7%	9.6%	11.7%	12.3%	23.3%
20	66	8.6%	7.6%	**16.7%**	**16.4%**	11.0%	13.3%	15.4%	24.8%
24	91	8.6%	7.2%	17.4%	**16.4%**	11.0%	13.7%	16.5%	25.3%

Table 1: Results on BANCA, SCface and MOBIO (EER on Dev set, HTER on Test set) showing the effect of block size during feature extraction ($B \times B$ pixel blocks with D DCT coefficients retained).

	BANCA		SCface	
	Dev	Test	Dev	Test
No score norm.	11.0%	11.1%	23.9%	25.1%
ZT-norm	**7.8%**	**6.1%**	**16.7%**	**16.4%**

Table 2: Results on BANCA and SCface (EER on Dev set, HTER on Test set) showing the effect of ZT-norm score normalisation.

System	Man. face loc.		Auto. face loc.	
	Dev	Test	Dev	Test
Baseline	7.8%	6.1%	9.2%	6.7%
ISV	**6.6%**	**5.4%**	**7.5%**	**6.0%**
JFA	7.6%	6.3%	9.1%	7.0%

Table 3: Results on BANCA (EER on Dev set, HTER on Test set) comparing different session variability modelling techniques, when using manual or automatic face localisation.

On both the development and test sets, the best performance was achieved using the ISV approach, which improved test set HTER by 11% relative. These improvements are statistically significant at a level of 95% and 85% for development and test sets respectively. Table 4 shows that using 50 to 100 dimensions in $\boldsymbol{U}$ was optimal on the development set, and this generalised well to the test set. It is encouraging that these results do not appear overly sensitive to the choice of subspace dimension.

Our results are compared to recently published work in Table 5. For comparison, we report the HTER on the test set ($g2$), development set ($g1$)[8], and the average. Note that Rua *et al*. [19] used automatic face extraction from the BANCA videos, while Ahonen *et al*. [1] used the 5 pre-selected images from each video as in this work. Our results represent a 31% reduction in average HTER when compared to previous work.

[8]This is obtained by applying the EER threshold from $g2$.

U dimensions	Man. face loc.		Auto. face loc.	
	Dev	Test	Dev	Test
None	7.8%	6.1%	9.2%	6.7%
10	7.0%	5.8%	8.1%	6.3%
50	**6.6%**	5.4%	8.1%	6.1%
100	6.7%	5.5%	**7.5%**	6.0%
150	6.7%	**5.3%**	7.7%	**5.4%**

Table 4: Results on BANCA (EER on Dev set, HTER on Test set) showing the effect of tuning the dimensionality of the session variability subspace, $\boldsymbol{U}$, when using manual or automatic face localisation.

System	Dev	Test	Average
Rúa *et al*. [19]	10.6%	9.8%	10.2%
Ahonen *et al*. [1]	-	-	9.1%
ISV (man.)	**7.1%**	**5.4%**	**6.3%**
ISV (auto.)	**7.7%**	**6.0%**	**6.8%**

Table 5: A comparison to previously published results (half total error rate) for the P protocol of the BANCA English database, on the Dev (g1) and Test (g2) sets, when using manual (man.) or automatic (auto.) face localisation.

5.3. Session variability modelling on SCface

On the SCface database, as shown in Table 6, both of the proposed session variability modelling techniques outperformed the baseline. JFA offered consistently improved performance over the baseline and ISV systems, resulting in a relative reduction in test set HTER of 18% over baseline results. The dimensionalities of $\boldsymbol{V}$ and $\boldsymbol{U}$ were tuned on the development set to values of 10 and 40, respectively. On the test set, the improvements provided by ISV and JFA

System	Dev	Test
Baseline	16.7%	16.4%
ISV	15.5%	14.3%
JFA	**12.0%**	**13.5%**

Table 6: Results on SCface (EER on Dev set, HTER on Test set) comparing different session variability modelling techniques.

System	Male		Female	
	Dev	Test	Dev	Test
Baseline	9.2%	10.5%	10.7%	20.4%
ISV	**4.0%**	8.3%	**6.1%**	**11.4%**
JFA	**4.0%**	**7.3%**	7.7%	13.0%

Table 7: Results on MOBIO, for males and females, comparing different session variability modelling techniques.

over the baseline are statistically significant at levels greater than 98% and 99% respectively.

In Table 8, results are further analysed by separating the scores into three separate groups, i.e. those from test images taken at the 3 different distances, *close*, *medium* and *far*. For this analysis only, the dataset used for Z-norm score normalisation was matched to the distance of the test image. Table 8 shows that the JFA approach provided substantial improvements for close and medium images, however, recognising far images remains particularly difficult.

5.4. Session variability modelling on MOBIO

On the MOBIO database, as shown in Table 7, both ISV and JFA substantially reduced the error rate compared to the baseline, with improvements statistically significant at a level greater than 99.99%. For the male tests, JFA outperformed ISV, providing a relative improvement over the baseline of 30%, using dimensionalities of 30 and 50 for $\boldsymbol{V}$ and $\boldsymbol{U}$ respectively. For female clients, the ISV technique performed the best, providing a relative improvement of 44% with 250 dimensions in $\boldsymbol{U}$.

6. Discussion

ISV consistently improved accuracy across all databases. JFA generally outperformed the baseline and was sometimes preferable to ISV, but not always. In particular, for BANCA, JFA was not helpful. In this case, we suspect that the world data of only 300 images of 30 clients was insufficient to accurately estimate $\boldsymbol{V}$ and $\boldsymbol{D}$. For MOBIO, much more training data was used. For ISV, this provided substantial improvements over the baseline, however, JFA did not outperform ISV for the female clients. In this case, we note that there was a significant gender imbalance in the MOBIO training data, with a male to female ratio of about 3:1, while the JFA training was conducted in a gender-independent manner. It is thus possible that the performance of JFA for female clients was disadvantaged by this imbalance. Therefore, future work should investigate gender-dependent subspace training for JFA, as well as training with additional data.

7. Conclusions

This work showed that session variability modelling can be used to improve face authentication accuracy. The techniques of inter-session variability modelling (ISV) and joint factor analysis (JFA), previously only applied to speaker authentication, were evaluated on several face authentication databases and were found to improve accuracy by up to 44% using limited training data. Our results on the BANCA database represent a 31% reduction in average HTER when compared to previous work. We found that ISV offered consistent improvements, while the results using JFA were less conclusive. In future work we plan to use additional training data to improve estimation of the subspaces, particularly for JFA, and also apply the proposed techniques to more databases. Further analysis may also give insights into the kind of information that is captured by the model, for example, the extent to which it captures describable sources of image variations.

References

[1] T. Ahonen and M. Pietikäinen. Pixelwise local binary pattern models of faces using kernel density estimation. In *Advances in Biometrics*, volume 5558 of *Lecture Notes in Computer Science*, pages 52–61. 2009. 1, 6

[2] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1-3):42–54, 2000. 2

[3] E. Bailly-Baillière et al. The BANCA database and evaluation protocol. In *Audio- and Video-Based Biometric Person Authentication*, volume 2688 of *Lecture Notes in Computer Science*, pages 1057–1071. 2003. 4

[4] S. Bengio and J. Mariéthoz. A statistical significance test for person authentication. In *Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop*, number Idiap-RR-83-2003, 2004. 5

[5] S. Bengio, J. Mariéthoz, and M. Keller. The expected performance curve. In *International Conference on Machine Learning, Workshop on ROC Analysis in Machine Learning*, 2005. 4

[6] F. Cardinaux, C. Sanderson, and S. Bengio. User authentication via adapted statistical models of face images. *IEEE Transactions on Signal Processing*, 54(1):361–373, 2006. 1, 2

System	Combined		Close		Medium		Far	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Baseline	16.7%	16.4%	13.6%	13.5%	14.1%	10.2%	18.6%	**19.7%**
JFA	**12.0%**	**13.5%**	**10.0%**	**11.2%**	**9.6%**	**8.0%**	**16.4%**	20.3%
	(-28%)	(-18%)	(-26%)	(-17%)	(-32%)	(-22%)	(-12%)	(+3%)

Table 8: Results on SCface protocols (EER on Dev set, HTER on Test set) showing the relative reduction in error rates when using joint factor analysis (JFA).

[7] F. Cardinaux, C. Sanderson, and S. Marcel. Comparison of MLP and GMM classifiers for face verification on XM2VTS. In *Audio- and Video-Based Biometric Person Authentication*, volume 2688 of *Lecture Notes in Computer Science*, pages 1058–1059. 2003. 1, 2

[8] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny. Comparison of scoring methods used in speaker recognition with joint factor analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4057–4060, 2009. 2, 3

[9] M. Grgic, K. Delac, and S. Grgic. SCface-surveillance cameras face database. *Multimedia tools and applications*, 51:863–879, 2011. 4, 5

[10] G. B. Huang, M. Ramesh, T. Berg, , and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 4

[11] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1435–1447, 2007. 1, 2, 3, 4

[12] S. Lucey and T. Chen. A GMM parts based face representation for improved verification through relevance adaptation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 855–861, 2004. 2

[13] C. McCool, V. Chandran, S. Sridharan, and C. Fookes. 3D face verification using a free-parts approach. *Pattern Recognition Letters*, 29:1190–1196, 2008. 2

[14] C. McCool and S. Marcel. Mobio database for the ICPR 2010 face and speech competition. Technical Report Idiap-Com-02-2009, Idiap Research Institute, 2009. 5

[15] M. McLaren, R. Vogt, B. Baker, and S. Sridharan. A comparison of session variability compensation approaches for speaker verification. *Information Forensics and Security, IEEE Transactions on*, 5(4):802–809, 2010. 1, 2

[16] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 947–954, 2005. 4

[17] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41, 2000. 1, 2, 3

[18] Y. Rodriguez. *Face detection and verification using local binary patterns*. PhD thesis, Idiap Research Institute and École Polytechnique Fédérale de Lausanne, 2006. 4

[19] E. Rúa, J. Castro, and C. Mateo. Quality-based score normalization for audiovisual person authentication. In *Image Analysis and Recognition*, volume 5112 of *Lecture Notes in Computer Science*, pages 1003–1012. 2008. 1, 6

[20] C. Sanderson and K. Paliwal. Fast features for face authentication under illumination direction changes. *Pattern Recognition Letters*, 24:2409–2419, 2003. 1, 2

[21] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Transactions on Image Processing*, 19(6):1635–1650, 2010. 2, 5

[22] O. Thyes, R. Kuhn, P. Nguyen, and J. Junqua. Speaker identification and verification using eigenvoices. In *ICSLP*, volume 2, pages 242–245, 2000. 4

[23] M. Turk and A. Pentland. Face recognition using eigenfaces. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591, 1991. 4

[24] R. Vogt and S. Sridharan. Explicit modelling of session variability for speaker verification. *Computer Speech & Language*, 22(1):17–38, 2008. 1, 2, 3, 5