IDIAP RESEARCH REPORT


# SESSION VARIABILITY MODELLING FOR FACE AUTHENTICATION

Chris McCool  Roy Wallace  Mitchell McLaren
Laurent El Shafey  Sébastien Marcel

Idiap-RR-17-2013

MAY 2013

**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

# Session Variability Modelling for Face Authentication

Christopher McCool[a,c,1,*], Roy Wallace[a], Mitchell McLaren[b], Laurent El Shafey[a], Sébastien Marcel[a]

[a]*Idiap Research Institute, Centre du Parc, 1920-Martigny, Switzerland.*
[b]*Radboud University Nijmegen, The Netherlands, PO Box 9102, 6500HC.*
[c]*NICTA, GPO Box 2434, Brisbane QLD 4001.*

**Abstract**

This paper examines session variability modelling for face authentication using Gaussian mixture models. Session variability modelling aims to explicitly model and suppress detrimental within-class (inter-session) variation. We examine two techniques to do this, inter-session variability modelling (ISV) and joint factor analysis (JFA), which were initially developed for speaker authentication. We present a self-contained description of these two techniques and demonstrate that they can be successfully applied to face authentication. In particular, we show that using ISV leads to significant error rate reductions of, on average, $26\%$ on the challenging and publicly-available databases SCface, BANCA, MOBIO, and Multi-PIE. Finally, we show that a limitation of both ISV and JFA for face authentication is that the session variability model captures and suppresses a significant portion of between-class variation.

*Keywords:* session variability modelling; joint factor analysis; Gaussian mixture

*Corresponding author

*Email addresses:* chris.mccool@nicta.com.au (Christopher McCool), roy.wallace@idiap.ch (Roy Wallace), m.mclaren@let.ru.nl (Mitchell McLaren), lefshafey@idiap.ch (Laurent El Shafey), marcel@idiap.ch (Sébastien Marcel)

[1]The majority of this work was completed while a post-doctoral researcher at the Idiap Research Institute.

## 1. Introduction

Many challenges in face authentication can be attributed to the problem of session variability. Session variability is anything that causes a mismatch between images of the same client and includes changes in illumination, pose, expression or image acquisition. Despite the fact that face authentication has evolved considerably in the past 15 years, modern approaches still suffer from increased errors in the presence of substantial session variability [1, 2] and this has been part of the motivation for the capture of recent face databases such as MOBIO [3] and Multi-PIE [4].

The same problem of session variability is faced in other fields such as speaker authentication, where it has been addressed by explicitly modelling the session variability. In speaker authentication, the detrimental session variation is caused by different microphones, acoustic environments and transmission channels. Two of the most successful techniques in improving robustness to session variability for speaker authentication are inter-session variability modelling (ISV) [5] and the related technique of joint factor analysis (JFA) [6], which have been shown to reduce errors by more than $30\%$. ISV and JFA aim to explicitly model and remove within-class (WC) variation using a low-dimensional subspace. JFA can be considered to be an extension of ISV as it additionally utilises a between-class (BC) subspace to capture important discriminative client information. ISV and JFA *explicitly model* session variation in that the set of observations from a particular biometric sample is modelled by a unique distribution over feature space, which is a function of a session-independent term plus a session-dependent term

that is *explicitly* estimated per-sample. These two techniques have been applied in the context of a Gaussian mixture model (GMM) based speaker authentication system [7].

A large variety of approaches have been proposed for face authentication. Most of them rely on a holistic representation of the face and make use of subspaces or manifolds, such as Kernel PCA Plus LDA [22] or Local Region PCA [21]. In addition, there is still healthy competition between approaches, as was shown recently in [23]. Inspired by speaker authentication techniques, a non-holistic GMM based face authentication system was proposed by Sanderson and Paliwal [8] which we refer to as the GMM parts-based approach. This GMM parts-based approach divides the face into blocks, treats each block independently and learns a GMM that describes all of these observations for a particular individual (client). In [9] and [10], different forms of relevance adaptation were proposed to improve the robustness of the approach and to better utilise limited enrolment data and an alternative set of features was proposed in [11]. In [10] it was found that the GMM parts-based approach offered the best trade-off in terms of complexity, robustness and discrimination, when compared with more complicated hidden Markov model (HMM) approaches. Also, the GMM parts-based approach forms the basis of the multiple region histogram (MRH) approach proposed in [12]. However, until recently explicitly modelling session variability was not considered as a way to improve the accuracy of this parts-based modelling approach for face authentication.

Our recent work [13] illustrated that ISV and JFA could be applied to a GMM-based face authentication framework. In this initial paper we showed that impressive relative performance improvements of up to $44\%$ could be obtained for the

GMM system, this yielded state-of-the-art performance on the BANCA database, with a relative improvement of $34\%$ when benchmarked against other techniques.

In this article we expand upon the initial work in [13], and hence focus on GMM parts-based approaches. There are four major contributions of this work[2]. First, we describe the ISV and JFA algorithms and provide all of the equations needed to implement them. To the best of our knowledge, this is the first time that these two algorithms have been presented in such a self-contained and easy to reproduce manner. Second, we extend the experiments to include the new Multi-PIE database. We do this so that we can train with many more identities to test the hypothesis of [13] that having more training identities will yield an improved JFA model; previously only 50 identities were available for training but with Multi-PIE we have 208 identities to train with. Third, we perform extensive comparisons against the related state-of-the-art MRH approach [12] and demonstrate experimentally that the GMM, ISV and JFA approaches all outperform the MRH approach. Fourth, we show through analysis that there is a limitation with both ISV and JFA which stems from the fact that, in practice, the session variability subspace captures and suppresses a significant amount of between-class (BC) variation. Finally, we suggest directions of future work to address this problem.

In Section 2 we describe the GMM-based system and then briefly describe the MRH approach. In Section 3, we describe the ISV and JFA techniques and provide all of the equations needed to implement these models. In Section 4 we outline our experimental protocol and present extended results in Section 5

[2]In addition, we found, highlighted and fixed a bug in the previous ISV results where the latent variable $\boldsymbol{z}$ was not being estimated while training the $\boldsymbol{U}$ subspace. While fixing this bug does not change the conclusions of the initial work, the new results in this article show slight improvements on most of the databases and thus supersede those in [13].

which includes a comparison with the MRH approach. In Section 6 we present our analysis of the ISV and JFA models and show that a significant amount of BC variation is removed by the session variability models. We then conclude and provide directions of future work in Section 7.

## 2. GMMs for face authentication

The GMM parts-based approach was first applied to face authentication in [8] and has since been successfully utilised by several researchers [9, 10]. This approach decomposes the face into a set of blocks that are considered to be separate observations of the same signal (the face) and it was found to offer the best trade-off in terms of complexity, robustness and discrimination [14, 10]. An overview of this procedure is given in Figure 1.

The key aspects for applying GMMs to face authentication are: (i) how to obtain the features (image pre-processing and feature extraction), (ii) how to make a model of each client (enrolment), and (iii) how to perform authentication given a probe image and a claimed client identity (testing). We deal with each of these below and then describe the state-of-the-art MRH system [12] which we use as a baseline system for our experiments.

### *2.1. Image pre-processing and feature extraction*

In this work, each image is rotated, cropped and registered to a $64 \times 80$ intensity image with the eyes 16 pixels from the top and separated by 33 pixels. Each cropped image is processed using Tan & Triggs pre-processing [15]; this was not applied to the images on SCface as their low resolution led to performance degradation when using this pre-processing technique. From each pre-processed image we exhaustively sample $B \times B$ blocks of pixel values by moving the sampling

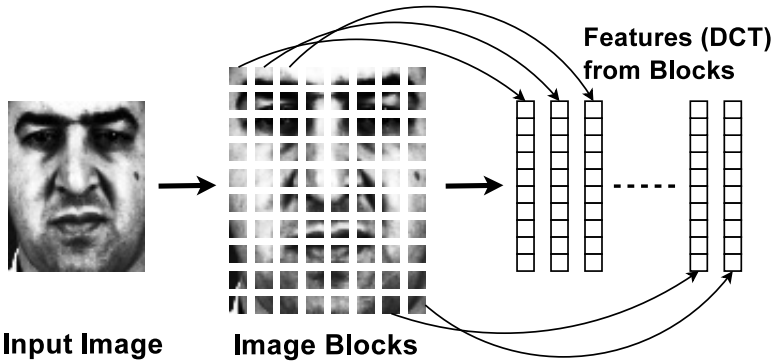



Figure 1: The concept of a parts-based approach: dividing the face into blocks and obtaining a feature vector from each block.

window one pixel at a time. Each block is mean and variance normalised prior to extracting the $M + 1$ lowest-frequency 2D discrete Cosine transform (2D-DCT) coefficients and then excluding the lowest frequency. The zeroth coefficient, or lowest frequency which is the offset, was removed as the pixel-based mean normalisation meant it was redundant. The resulting 2D-DCT feature vectors are mean and variance normalised in each dimension, with respect to the other feature vectors of the image. Each image is thus represented by a set of $K$ feature vectors, $\boldsymbol{O} = \left\{\boldsymbol{o}^1, \boldsymbol{o}^2, \ldots, \boldsymbol{o}^K\right\}$, each of dimensionality $M$.

### *2.2. Creating a client model*

A GMM is a generative model comprised of $C$ Gaussian components [7]. Each component is defined by its weight, $\omega_c$, mean vector, $\boldsymbol{\mu}_c$, and covariance matrix (considered to be diagonal in our work), $\boldsymbol{\Sigma}_c$, such that,

$$Pr\left(\boldsymbol{O}|\boldsymbol{\Theta}\right) = \prod_{k=1}^{K} \sum_{c=1}^{C} \omega_c \mathcal{N}\left[\boldsymbol{o}^k|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\right]. \tag{1}$$

The parameters, $\boldsymbol{\Theta}$, of this model consist of the $C$ weights, means and variances. Given an image $\boldsymbol{O}_t$, the zeroth and first order statistics for the $c^{th}$ component of this model are,

$$n_{t;c} = \sum_{k=1}^{K} \gamma_c(\boldsymbol{o}_t^k) \quad \text{and} \quad \boldsymbol{f}_{t;c} = \sum_{k=1}^{K} \gamma_c(\boldsymbol{o}_t^k)\boldsymbol{o}_t^k, \tag{2}$$

respectively, where the term $\gamma_c(\boldsymbol{o}_t^k)$ is the occupation probability of the $k^{th}$ observation of image $t$ for the $c^{th}$ component and is given by

$$\gamma_c(\boldsymbol{o}_t^k) = \frac{\omega_c \mathcal{N}\left[\boldsymbol{o}_t^k | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\right]}{\sum_{c=1}^{C} \omega_c \mathcal{N}\left[\boldsymbol{o}_t^k | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\right]}. \tag{3}$$

In the GMM parts-based approach to face authentication the distribution of local features from images of a person's face is modelled by a GMM. To use GMMs for authentication we thus need to be able to generate a model for each client $i$ given a set of enrolment images. The main difficulty with doing this is that typically we have a limited number of enrolment samples per client. To overcome this difficulty we use a universal background model (UBM), or world model, as a prior and adapt this prior to better match the enrolment samples of a client. This is achieved by using maximum *a posteriori* (MAP) estimation and was termed relevance MAP adaptation in [7]. This allows us to generate a client model $\boldsymbol{s}_i$ with limited amounts of training data and in practice it has been shown that mean-only adaptation, where only the means of the UBM are adapted, is effective for speaker [7] and face authentication [9, 10].

Mean-only relevance MAP can be written in a compact way by using GMM super-vector notation, this will also provide a compact representation for the session variability modelling techniques ISV and JFA. Super-vector notation consists of taking the parameters (weights, means and covariance matrices) of a GMM and creating a single vector or matrix to represent each of them. An example of this would be that the means of the UBM can be concatenated to form a single mean super-vector given by $\boldsymbol{m} = [\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T, \ldots, \boldsymbol{\mu}_C^T]^T$. Using this notation it was shown

in [5] that mean-only relevance MAP adaptation equates to being,

$$\boldsymbol{s}_i = \boldsymbol{m} + \boldsymbol{d}_i, \tag{4}$$

where the client model is given by $\boldsymbol{s}_i$ which is a mean super-vector consisting of two parts: (i) the prior world model, $\boldsymbol{m}$, and (ii) a client-specific offset $\boldsymbol{d}_i$. The client-specific offset $\boldsymbol{d}_i$ is

$$\boldsymbol{d}_i = \boldsymbol{D}\boldsymbol{z}_i, \tag{5}$$

where $\boldsymbol{D}$ is a diagonal matrix of size $(CM, CM)$ given by

$$\boldsymbol{D} = \sqrt{\frac{\boldsymbol{\Sigma}}{\tau}} \tag{6}$$

and $\boldsymbol{\Sigma}$ is a block diagonal matrix with block diagonal entries consisting of the covariance matrices $\boldsymbol{\Sigma}_c$ for each of the $C$ components of the UBM. The term $\tau$ is a relevance factor and provides a weight for the prior when performing MAP adaptation [7]. The latent variable $\boldsymbol{z}_i$ is assumed to be normally distributed, $\boldsymbol{z}_i \sim \mathcal{N}(0, \boldsymbol{I})$.

Creating a client model is achieved by finding the MAP solution of $\boldsymbol{z}_i$ which is given by,

$$\boldsymbol{z}_i = \left(\tau \boldsymbol{I} + \boldsymbol{N}_i\right)^{-1} \boldsymbol{f}_{i|\boldsymbol{m}}. \tag{7}$$

The terms $\boldsymbol{N}_i$ and $\boldsymbol{f}_{i|\boldsymbol{m}}$ refer to the zeroth order and mean centralised first order statistics of the $J_i$ enrolment images of the $i^{th}$ client. The mean centralised first order statistic is

$$\boldsymbol{f}_{i|\boldsymbol{m}} = \sum_{j=i}^{J_i} \boldsymbol{f}_{i,j|\boldsymbol{m}} \tag{8}$$

where the mean centralised first order statistic for the $j^{th}$ image of client $i$ is

$$\boldsymbol{f}_{i,j|\boldsymbol{m}} = \boldsymbol{f}_{i,j} - \boldsymbol{N}_{i,j}\boldsymbol{m}, \tag{9}$$

$\boldsymbol{f}_{i,j} = [\boldsymbol{f}_{i,j;1}^T, \boldsymbol{f}_{i,j;2}^T, \ldots, \boldsymbol{f}_{i,j;C}^T]^T$, and $\boldsymbol{f}_{i,j;c}$ is the first order statistic of component

$c$ for the $j^{th}$ image of client $i$ (2). The zeroth order statistic of client $i$ is,

$$\boldsymbol{N}_i = \sum_{j=1}^{J_i} \boldsymbol{N}_{i,j}, \tag{10}$$

where

$$\boldsymbol{N}_{i,j} = \begin{bmatrix} \boldsymbol{N}_{i,j;1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{N}_{i,j;C} \end{bmatrix}, \quad \text{and} \quad \boldsymbol{N}_{i,j;c} = \begin{bmatrix} n_{i,j;c} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & n_{i,j;c} \end{bmatrix}. \tag{11}$$

The term $n_{i,j;c}$ is the zeroth order statistic of component $c$ for the $j^{th}$ image of client $i$ (2) and $\boldsymbol{N}_{i,j;c}$ is of size $(M, M)$.

### *2.3. Classification*

At test time we need to make a decision as to whether or not the features extracted from a test image, $\boldsymbol{O}_t$, were generated by the model of the claimed client identity ($\boldsymbol{s}_i$). To do this we use a log-likelihood ratio (LLR) [7] to compare the hypotheses that $\boldsymbol{O}_t$ was generated by the client's model $\boldsymbol{s}_i$ versus the hypothesis that it was not generated by the client. For the second, negative hypothesis, we evaluate the log-likelihood from the UBM, $\boldsymbol{m}$, and so the LLR becomes,

$$h\left(\boldsymbol{O}_t, \boldsymbol{s}_i\right) = \sum_{k=1}^{K} \left(\ln(p(\boldsymbol{o}_t^k \mid \boldsymbol{s}_i)) - \ln(p(\boldsymbol{o}_t^k \mid \boldsymbol{m}))\right). \tag{12}$$

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, $\theta$.

In this work, we use a fast scoring technique known as *linear scoring* [16], which is an approximation of the log-likelihood ratio that was shown to be as accurate and up to two orders of magnitude more efficient to compute. In GMM super-vector notation linear scoring can be simply written as,

$$h_{\text{linear}}\left(\boldsymbol{O}_t, \boldsymbol{s}_i\right) = \left(\boldsymbol{s}_i - \boldsymbol{m}\right)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{f}_{t|\boldsymbol{m}}. \tag{13}$$

*2.4. Baseline MRH system*

The state-of-the-art MRH approach [12] is closely related to the GMM mean-only MAP approach. The MRH approach creates a probabilistic bag of visual words by using a GMM UBM. As with the standard GMM approach the face is divided into a set of blocks and from each block a set of DCT features are obtained, however, rather than perform a mean-only MAP adaptation a probabilistic histogram is formed. In our notation this probabilistic histogram is equivalent to forming a normalised feature vector of the zeroth order statistics (occupation probabilities) (2) given the image $\boldsymbol{O}_t$,

$$\boldsymbol{l} = \left[n_{t,1}, n_{t,2}, \ldots, n_{t,C}\right]^T / N_t, \tag{14}$$

where $N_t$ is the number of DCT features and ensures that $\boldsymbol{l}$ sums to 1.

This is a multi-region technique as it obtains a probabilistic histogram $\boldsymbol{l}_r$ for a set of $r = [1, \ldots, R]$ pre-defined regions. It was found that using a $3 \times 3$ grid to define $R = 9$ regions provided good performance [12]. At enrolment time a model for the $i^{th}$ client $\boldsymbol{s}_i$ is formed by obtaining the probabilistic histogram $\boldsymbol{l}_{i,r}$ using all of the samples for the $r^{th}$ region from all of the enrolment images. At test time, the probabilistic histogram for each region of the test image $\boldsymbol{O}_t$, $\boldsymbol{l}_{t,r}$, is compared against the probabilistic histogram of the same region for the client, $\boldsymbol{l}_{i,r}$, using an L1 distance. These distances are then summed to obtain a score,

$$h_{L1}\left(\boldsymbol{O}_t, \boldsymbol{s}_i\right) = \sum_{r=1}^{R} \|\boldsymbol{l}_{i,r}, \boldsymbol{l}_{t,r}\|_1. \tag{15}$$

The image $\boldsymbol{O}_t$ is then classified as belonging to client $i$ if and only if $h_{L1}\left(\boldsymbol{O}_t, \boldsymbol{s}_i\right)$ is greater than a threshold, $\theta$.

This probabilistic histogram approach can be viewed as a simplification of GMM mean-only MAP adaptation since it uses only the zeroth order statistics to describe the client and test sample. By comparison the GMM mean-only MAP approach incorporates both the zeroth and first order statistics. Finally, the probabilistic histogram approach compares the feature vectors using an $L1$ distance whereas the GMM baseline uses the linear scoring approximation of the log-likelihood which equates to a Mahalanobis distance (13).

## 3. Session variability modelling

Inter-session variability modelling (ISV) [5] and joint factor analysis (JFA) [6] are two *session variability modelling* techniques that have been applied with success to speaker authentication. This section provides an overview of these techniques and how they can be applied to face authentication. In the context of face authentication, session variability refers to the variability that results in differences between images of the same person. For instance, pose and illumination variation can result in significantly different images even though the identity remains constant. Some examples of this session variability can be seen in Figure 4.

ISV and JFA are applied in the context of a GMM-based system. In the case of mean-only relevance MAP adaptation, as described in Section 2, there is no explicit modelling of session variability and so the model consists of only two parts, the UBM ($\boldsymbol{m}$) and the client-specific offset ($\boldsymbol{d}_i$) as described by (4). 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 (4), and so this will likely lead to a sub-optimal 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 compensate for this variation during enrolment as well as testing. This is achieved by excluding sources of session variation when generating a client's model as well as estimating and compensating for the different conditions (session variations) observed in test images. This approach of session variability modelling is highly advantageous as it can be used in conjunction with state-of-the-art image normalisation techniques to model the residual noise which will inevitably be left behind; no normalisation technique is perfect and so there will always be some residual form of noise or session variation.

When we apply session variability modelling to face authentication we consider that each image corresponds to a different session. This seems intuitive because each image can be captured with a different facial expression, pose or even illumination. Following [5], the particular conditions of a session are assumed to result in an additive offset to $\boldsymbol{s}_i$ which can be expressed as

$$\boldsymbol{\mu}_{i,j} = \boldsymbol{s}_i + \boldsymbol{u}_{i,j}, \tag{16}$$

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 super-vector 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 $\boldsymbol{u}_{i,j}$.

Both ISV and JFA *explicitly model* session variability, however, JFA also explicitly models between-client variability. This difference, along with the algorithms used for estimation, training and classification, will be discussed in more detail in the following sections. This is the first time that such a self-contained

description of these algorithms has been provided for face authentication.

### *3.1. Inter-session variability modelling (ISV)*

The ISV technique, proposed in [5], assumes that within-client variation is contained in a linear subspace of the GMM mean super-vector space. That is,

$$\boldsymbol{u}_{i,j} = \boldsymbol{U}\boldsymbol{x}_{i,j}, \tag{17}$$

where $\boldsymbol{U}$ is the low-dimensional subspace of size $(CM, n_x)$ that contains within-client variation, and $\boldsymbol{x}_{i,j}$, of size $(n_x, 1)$, is the latent session variable which is assumed to be normally distributed ($\boldsymbol{x}_{i,j} \sim \mathcal{N}(0, \boldsymbol{I})$). As with relevance MAP adaptation, the client-dependent offset is set to $\boldsymbol{d}_i = \boldsymbol{D}\boldsymbol{z}_i$, as per (5) and (6).

To summarise, in this generative model each image is assumed to have been generated by a GMM mean super-vector

$$\boldsymbol{\mu}_{i,j} = \boldsymbol{m} + \boldsymbol{U}\boldsymbol{x}_{i,j} + \boldsymbol{D}\boldsymbol{z}_i. \tag{18}$$

At enrolment time, the model for client $i$ is obtained by estimating the latent variables, $\boldsymbol{z}_i$ and $\boldsymbol{x}_{i,j}$, using the procedure described in Section 3.3. The estimated effect of session variability in each image (17) is then excluded from the client model. This means that for ISV the resulting client model is

$$\boldsymbol{s}_i^{ISV} = \boldsymbol{m} + \boldsymbol{D}\boldsymbol{z}_i. \tag{19}$$

This should not be confused with relevance MAP adaptation (4) because to obtain $\boldsymbol{s}_i^{ISV}$ the latent identity variable $\boldsymbol{z}_i$ is estimated along with the latent session variable $\boldsymbol{x}_{i,j}$ in the generative framework defined by (18), thus suppressing the effects of session variability and so the client model for ISV, $\boldsymbol{s}_i^{ISV}$, will be quite different to the one for relevance MAP adaptation, $\boldsymbol{s}_i$. Scoring for ISV is discussed in Section 3.5.

*3.2. Joint factor analysis (JFA)*

JFA [6] can be seen as an extension of ISV. Specifically, for JFA the client-dependent offset is defined as

$$\boldsymbol{d}_i = \boldsymbol{V}\boldsymbol{y}_i + \hat{\boldsymbol{D}}\boldsymbol{z}_i, \tag{20}$$

in contrast to relevance MAP adaptation and ISV where $\boldsymbol{d}_i = \boldsymbol{D}\boldsymbol{z}_i$. For JFA, $\boldsymbol{V}$ is a low rank rectangular matrix of size $(CM, n_y)$, $\boldsymbol{y}_i$ is the latent identity variable of size $(n_y, 1)$ which is assumed to be normally distributed ($\boldsymbol{y}_i \sim \mathcal{N}(0, \boldsymbol{I})$), and $\boldsymbol{d}_i$ is thus distributed with covariance matrix $\hat{\boldsymbol{D}}^2 + \boldsymbol{V}\boldsymbol{V}^\top$. 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 modelling technique [17]. One of the motivations for using JFA is to improve enrolment of a client with limited data, by allowing a client model to be approximately represented by only the small number of factors in the latent identity variable $\boldsymbol{y}_i$.

To summarise, in contrast to ISV (18), for JFA each image is modelled by

$$\boldsymbol{\mu}_{i,j} = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y}_i + \boldsymbol{U}\boldsymbol{x}_{i,j} + \hat{\boldsymbol{D}}\boldsymbol{z}_i. \tag{21}$$

In this case, both $\boldsymbol{V}$ and $\hat{\boldsymbol{D}}$ are learnt from training data, in addition to $\boldsymbol{U}$, using maximum likelihood [6] (see Section 3.4). At enrolment time, the model for client $i$ is obtained by estimating the latent variables $\boldsymbol{x}_{i,j}$, $\boldsymbol{y}_i$ and $\boldsymbol{z}_i$ (see Section 3.3). As with ISV, we then suppress the effects of session variability by removing the term $\boldsymbol{U}\boldsymbol{x}_{i,j}$. For JFA, the resulting client model is

$$\boldsymbol{s}_i^{JFA} = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y}_i + \hat{\boldsymbol{D}}\boldsymbol{z}_i. \tag{22}$$

Scoring is similar to ISV and is discussed in Section 3.5.

In order to use the ISV and JFA frameworks described above we need to be able to: (i) estimate the latent variables, $\boldsymbol{x}_{i,j}$, $\boldsymbol{y}_i$ and $\boldsymbol{z}_i$, and (ii) train the subspaces

$\boldsymbol{U}$, $\boldsymbol{V}$ and $\hat{\boldsymbol{D}}$. As described below, to solve these two problems we follow the approach of [5] for ISV which is, in short, MAP estimation to solve problem (i) and maximum likelihood (ML) estimation to solve problem (ii).

### *3.3. Estimation of latent variables*

The approach to estimating the latent variables is the same for ISV and JFA. The aim is to jointly estimate the latent variables, $\boldsymbol{x}_{i,j}$, $\boldsymbol{y}_i$ and $\boldsymbol{z}_i$, using MAP estimation. In the case of ISV only $\boldsymbol{x}_{i,j}$ and $\boldsymbol{z}_i$ need to be estimated. These latent variables are of size $(n_x, 1)$, $(n_y, 1)$ and $(CM, 1)$ for $\boldsymbol{x}_{i,j}$, $\boldsymbol{y}_i$ and $\boldsymbol{z}_i$ respectively.

Central to this process is to note that the latent identity variables ($\boldsymbol{z}_i$ for ISV and additionally $\boldsymbol{y}_i$ for JFA) are tied together for all of the $J_i$ enrolment images of client $i$. This means that all $J_i$ enrolment images share the same latent identity variables but have different latent session variables $[\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, \ldots, \boldsymbol{x}_{i,J_i}]$. We can represent this in a convenient way by the set of equations

$$\begin{bmatrix} \boldsymbol{\mu}_{i,1} \\ \vdots \\ \boldsymbol{\mu}_{i,j} \end{bmatrix} = \begin{bmatrix} \boldsymbol{m} \\ \vdots \\ \boldsymbol{m} \end{bmatrix} + \tilde{\boldsymbol{A}}\tilde{\boldsymbol{\lambda}}_i, \tag{23}$$

where we have concatenated the latent variables of the client $i$ to form, for JFA,

$$\tilde{\boldsymbol{\lambda}}_i = [\boldsymbol{z}_i^T, \boldsymbol{y}_i^T, \boldsymbol{x}_{i,1}^T, \boldsymbol{x}_{i,2}^T, \ldots, \boldsymbol{x}_{i,J_i}^T]^T, \tag{24}$$

and $\tilde{\boldsymbol{A}}$ is a composite matrix with $J_i$ entries of $\boldsymbol{U}$, $\hat{\boldsymbol{D}}$ and $\boldsymbol{V}$, with $\boldsymbol{U}$ being repeated in a block diagonal fashion such that,

$$\tilde{\boldsymbol{A}} = \begin{bmatrix} \hat{\boldsymbol{D}} & \boldsymbol{V} & \boldsymbol{U} & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \vdots & \boldsymbol{0} & \ddots & \boldsymbol{0} \\ \hat{\boldsymbol{D}} & \boldsymbol{V} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{U} \end{bmatrix}. \tag{25}$$

We use a similar approach to represent ISV by simply removing the columns refer-

ring to $\boldsymbol{V}$ in (25) and its associated latent variable $\boldsymbol{y}_i$ in (24); also note that for ISV $\boldsymbol{D}$ is defined by (6) rather than $\hat{\boldsymbol{D}}$ which is learnt from data. For clarity we note that the size of the matrices involved are the following: $\hat{\boldsymbol{D}}$ is assumed diagonal and is of size $(CM, CM)$, $\boldsymbol{V}$ is rectangular of size $(CM, n_y)$ and $\boldsymbol{U}$ is rectangular of size $(CM, n_x)$. Thus the matrix $\tilde{\boldsymbol{A}}$ is of size $(J_i \times CM, CM + n_y + J_i \times n_x)$ and $\tilde{\boldsymbol{\lambda}}_i$ is a vector of size $(CM + n_y + J_i \times n_x, 1)$.

Using the above formulation, client enrolment reduces to finding the MAP estimate of $\tilde{\boldsymbol{\lambda}}_i$,

$$\begin{aligned}\tilde{\boldsymbol{\lambda}}_i^* &= \underset{\tilde{\boldsymbol{\lambda}}_i}{\operatorname{argmax}} \;\; p(\tilde{\boldsymbol{\lambda}}_i \mid \boldsymbol{O}_{i,1}, \boldsymbol{O}_{i,2}, \ldots, \boldsymbol{O}_{i,J_i}), \\ &= \underset{\tilde{\boldsymbol{\lambda}}_i}{\operatorname{argmax}} \;\; p(\boldsymbol{z}_i) p(\boldsymbol{y}_i) \prod_{j=1}^{J_i} p(\boldsymbol{O}_{i,j} \mid \boldsymbol{x}_{i,j}, \boldsymbol{y}_i, \boldsymbol{z}_i) p(\boldsymbol{x}_{i,j}). \end{aligned} \tag{26}$$

where $\boldsymbol{y}_i$ is omitted in the case of ISV. Solving this leads to the solution [18],

$$\tilde{\boldsymbol{\lambda}}_i^* = E\left[\tilde{\boldsymbol{\lambda}}_i\right] = \left(\boldsymbol{I} + \tilde{\boldsymbol{A}}^T \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{N}}_i \tilde{\boldsymbol{A}}\right)^{-1} \tilde{\boldsymbol{A}}^T \tilde{\boldsymbol{\Sigma}}^{-1} \left(\sum_{i=1}^{J_i} \boldsymbol{f}_{i,j|\boldsymbol{m}}\right), \tag{27}$$

where

$$\tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Sigma} \end{bmatrix} \quad \text{and} \quad \tilde{\boldsymbol{N}}_i = \begin{bmatrix} \boldsymbol{N}_{i,1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{N}_{i,J_i} \end{bmatrix}. \tag{28}$$

To solve (27) we use a Gauss-Seidel approximation method inspired from [5]. This approximation method was proposed for ISV in [5] and is necessary because $\tilde{\boldsymbol{A}}$ grows quadratically with respect to the number of images $J_i$ and so inverting the matrix $\left(\boldsymbol{I} + \tilde{\boldsymbol{A}}^T \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{N}}_i \tilde{\boldsymbol{A}}\right)$ quickly becomes intractable, even if we try to exploit its structure (see Section 3.3.2 of [5] for more details).

The Gauss-Seidel algorithm, Algorithm 1, iteratively estimates each latent

**Algorithm 1** Estimating Latent Variables for Identity $i$

1: $\boldsymbol{y}_i = \mathbf{0}$, $\boldsymbol{z}_i = \mathbf{0}$ and $\boldsymbol{x}_{i,j} = \mathbf{0}$; $j = 1, \ldots, J_i$
2: Estimate $\boldsymbol{N}_{i,j}$ and $\boldsymbol{f}_{i,j}$; $j = 1, \ldots, J_i$
3: $\boldsymbol{N}_i = \sum_{j=1}^{J_i} \boldsymbol{N}_{i,j}$
4: $\boldsymbol{f}_{i|\boldsymbol{m}} = \sum_{j=1}^{J_i} \boldsymbol{f}_{i,j|\boldsymbol{m}}$
5: **for** $gs = 1$ to Number of Gauss-Seidel iterations **do**
6: **Estimate** $E\left[\boldsymbol{y}_i\right]$
7: **for** $j = 1$ to $J_i$ **do**
8: **Estimate** $E\left[\boldsymbol{x}_{i,j}\right]$
9: **end for**
10: **Estimate** $E\left[\boldsymbol{z}_i\right]$
11: **end for**
12: **return** $E\left[\boldsymbol{y}_i\right]$, $E\left[\boldsymbol{z}_i\right]$, $\left[E\left[\boldsymbol{x}_{i,1}\right], \ldots, E\left[\boldsymbol{x}_{i,J_i}\right]\right]$

variable. It does this by factorising the concatenated latent variable $\tilde{\boldsymbol{\lambda}}_i$ into its respective latent variables $\boldsymbol{z}_i$, $\boldsymbol{y}_i$, $\boldsymbol{x}_{i,1}$ through to $\boldsymbol{x}_{i,J_i}$, this takes advantage of the known structure for these latent variables. Each factorised latent variable is then estimated using the most recent estimate of all of the other latent variables, doing this means that we no longer jointly estimate each latent variable but estimate a latent variable by considering all of the others to be fixed (or known). This simplifies the estimation steps as we now only need to solve for one latent variable, a more detailed description and motivation for this approach is given in Section 3.4 of [5]. We initialise this algorithm by setting all of the latent variables to $\mathbf{0}$, as they are assumed to be $\mathcal{N}\left(\mathbf{0}, \boldsymbol{I}\right)$. For the case of ISV we omit the step of estimating $E\left[\boldsymbol{y}_i\right]$ (line 6 of Algorithm 1) and $\boldsymbol{y}_i$ is effectively set to $\mathbf{0}$ in (29) and (31). The MAP estimation of each latent variable is,

$$E\left[\boldsymbol{x}_{i,j}\right] = \left(\boldsymbol{I} + \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{N}_{i,j}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\left[\boldsymbol{f}_{i,j|m} - \boldsymbol{N}_{i,j}(\boldsymbol{V}\boldsymbol{y}_i + \boldsymbol{D}\boldsymbol{z}_i)\right], \quad (29)$$

$$E\left[\boldsymbol{y}_i\right] = \left(\boldsymbol{I} + \boldsymbol{V}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{N}_i\boldsymbol{V}\right)^{-1}\boldsymbol{V}^T\boldsymbol{\Sigma}^{-1}\left[\boldsymbol{f}_{i|m} - \boldsymbol{N}_i\boldsymbol{D}\boldsymbol{z}_i - \sum_{j=1}^{J_i}\boldsymbol{N}_{i,j}\boldsymbol{U}\boldsymbol{x}_{i,j}\right], \tag{30}$$

$$E\left[\boldsymbol{z}_i\right] = \left(\boldsymbol{I} + \boldsymbol{D}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{N}_i\boldsymbol{D}\right)^{-1}\boldsymbol{D}^T\boldsymbol{\Sigma}^{-1}\left[\boldsymbol{f}_{i|m} - \boldsymbol{N}_i\boldsymbol{V}\boldsymbol{y}_i - \sum_{j=1}^{J_i}\boldsymbol{N}_{i,j}\boldsymbol{U}\boldsymbol{x}_{i,j}\right]. \tag{31}$$

### *3.4. Estimation of subspaces*

To learn the subspaces $\hat{\boldsymbol{D}}$, $\boldsymbol{V}$ and $\boldsymbol{U}$ we use an expectation-maximisation (EM) algorithm similar to that described in Section 5.2 of [5] for ISV. This algorithm consists of an expectation step where MAP estimates of the latent variables are made (see Section 3.3) and a maximisation step where the parameters are updated using ML. For JFA we learn $\boldsymbol{V}$, $\boldsymbol{U}$ and $\hat{\boldsymbol{D}}$ while for ISV we only learn $\boldsymbol{U}$. It can be shown that the updates for the parameters ($\boldsymbol{V}$, $\boldsymbol{U}$ and $\hat{\boldsymbol{D}}$) are obtained by solving the following systems of equations,

$$\begin{aligned}\boldsymbol{V}_c\left(\sum_{i=1}^{I}\boldsymbol{N}_{i;c}E\left[\boldsymbol{y}_i\boldsymbol{y}_i^T\right]\right) = \\ \sum_{i=1}^{I}\left(\sum_{j=1}^{J_i}\left(\boldsymbol{f}_{i,j;c|\boldsymbol{m}} - \boldsymbol{N}_{i,j;c}(\hat{\boldsymbol{D}}_c\boldsymbol{z}_i + \boldsymbol{U}_c\boldsymbol{x}_{i,j})\right)\right)E\left[\boldsymbol{y}_i\right]^T,\end{aligned} \tag{32}$$

$$\begin{aligned}\boldsymbol{U}_c\left(\sum_{i=1}^{I}\sum_{j=1}^{J_i}\boldsymbol{N}_{i,j;c}E\left[\boldsymbol{x}_{i,j}\boldsymbol{x}_{i,j}^T\right]\right) = \\ \sum_{i=1}^{I}\left(\sum_{j=1}^{J_i}\left(\boldsymbol{f}_{i,j;c|\boldsymbol{m}} - \boldsymbol{N}_{i,j;c}(\boldsymbol{V}_c\boldsymbol{y}_i + \hat{\boldsymbol{D}}_c\boldsymbol{z}_i)\right)E\left[\boldsymbol{x}_{i,j}\right]^T\right),\end{aligned} \tag{33}$$

$$\hat{\boldsymbol{D}}_c \left( \sum_{i=1}^{I} \boldsymbol{N}_{i;c} E\left[\boldsymbol{z}_i \boldsymbol{z}_i^T\right] \right) = \sum_{i=1}^{I} \left( \sum_{j=1}^{J_i} \left( \boldsymbol{f}_{i,j;c|\boldsymbol{m}} - \boldsymbol{N}_{i,j;c} (\boldsymbol{V}_c \boldsymbol{y}_i + \boldsymbol{U}_c \boldsymbol{x}_{i,j}) \right) \right) E\left[\boldsymbol{z}_i\right]^T . \tag{34}$$

Where $\boldsymbol{f}_{i,j;c|\boldsymbol{m}}$ is the mean normalised first order statistics for component $c$ similar to (9),

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}_1 \\ \vdots \\ \boldsymbol{V}_C \end{bmatrix}, \quad \boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}_1 \\ \vdots \\ \boldsymbol{U}_C \end{bmatrix}, \quad \text{and} \quad \hat{\boldsymbol{D}} = \begin{bmatrix} \hat{\boldsymbol{D}}_1 & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \hat{\boldsymbol{D}}_C \end{bmatrix}. \tag{35}$$

In this formulation, the expected value of the square of the latent variables is given by,

$$E\left[\boldsymbol{y}_i \boldsymbol{y}_i^T\right] = \left(\boldsymbol{I} + \boldsymbol{V}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{N}_i \boldsymbol{V}\right)^{-1} + E\left[\boldsymbol{y}_i\right] E\left[\boldsymbol{y}_i\right]^T , \tag{36}$$

$$E\left[\boldsymbol{x}_{i,j} \boldsymbol{x}_{i,j}^T\right] = \left(\boldsymbol{I} + \boldsymbol{U}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{N}_{i,j} \boldsymbol{U}\right)^{-1} + E\left[\boldsymbol{x}_{i,j}\right] E\left[\boldsymbol{x}_{i,j}\right]^T , \tag{37}$$

$$E\left[\boldsymbol{z}_i \boldsymbol{z}_i^T\right] = \left(\boldsymbol{I} + \hat{\boldsymbol{D}}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{N}_i \hat{\boldsymbol{D}}\right)^{-1} + E\left[\boldsymbol{z}_i\right] E\left[\boldsymbol{z}_i\right]^T . \tag{38}$$

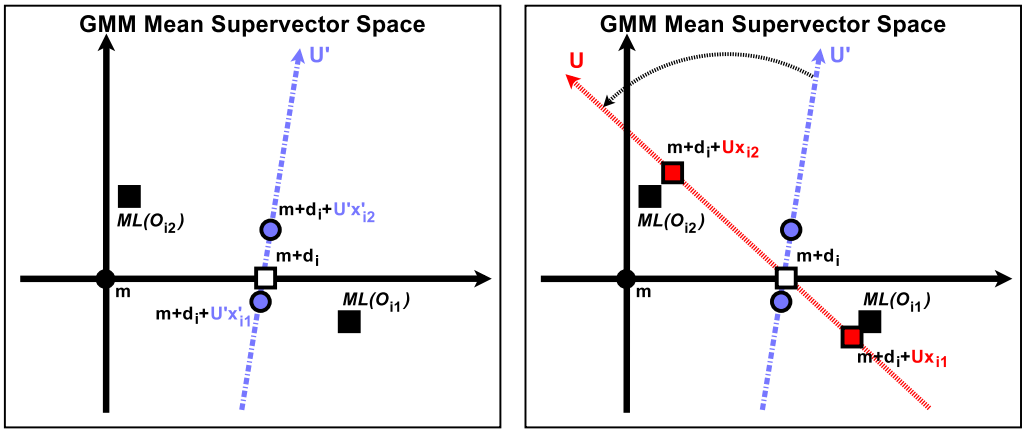
Note that for ISV we need to substitute the learnt matrix $\hat{\boldsymbol{D}}$ for the pre-defined matrix $\boldsymbol{D}$ in (33), (35) and (38).

We use the training procedure provided in the JFA cookbook, which is described in [19] and is similar to the one proposed in [18]. For JFA, this procedure first learns $\boldsymbol{V}$, then $\boldsymbol{U}$ and finally $\hat{\boldsymbol{D}}$. Each parameter is learnt using an EM algorithm, where the E-Step is the same as the one described in Section 3.3 and the M-Step is given by the equations above; those matrices which have yet to be learnt

are set to $\mathbf{0}$.

An illustration of how this process works is given in Figure 2 for training $\boldsymbol{U}$. In these illustrations we show how the session subspace $\boldsymbol{U}$ is updated. We have represented the training images in GMM mean super-vector space as maximum likelihood points for illustration purposes only. These illustrative points are equivalent to the maximum likelihood GMM mean super-vector for $\boldsymbol{O}_t$,

$$ML(\boldsymbol{O}_t) = \underset{\boldsymbol{\mu}}{\operatorname{argmax}}\, p(\boldsymbol{O}_t \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\omega}). \tag{39}$$




(a) Using an initial estimate $\boldsymbol{U}'$. (b) Updating the estimate $\boldsymbol{U}'$ to be $\boldsymbol{U}$.

Figure 2: This is an illustration of one step of the EM training algorithm for updating the $\boldsymbol{U}$ matrix in GMM mean super-vector space. In (a) and (b), $\boldsymbol{m} + \boldsymbol{d}_i$ is assumed to be known for the client $i$. We also provide, for illustrative purposes, the maximum likelihood (ML) points, for two images $ML(\boldsymbol{O}_{i1})$ and $ML(\boldsymbol{O}_{i2})$, which are given by (39). In (a) an initial estimate $\boldsymbol{U}'$ is given along with the associated maximum *a posteriori* (MAP) estimates of the session dependent offsets which would result in the illustrated blue circles given by $(\boldsymbol{m}+\boldsymbol{d}_i+\boldsymbol{U}'\boldsymbol{x}'_{i1})$ and $(\boldsymbol{m}+\boldsymbol{d}_i+\boldsymbol{U}'\boldsymbol{x}'_{i2})$; we have used the $\boldsymbol{U}'$ and $\boldsymbol{x}'$ to indicate the matrix and associated latent variables during the first EM iteration. In (b) we show how the initial estimate $\boldsymbol{U}'$ may be updated to $\boldsymbol{U}$ such that the likelihood of the training data is increased, with the new MAP estimates illustrated by red squares, and given by $(\boldsymbol{m} + \boldsymbol{d}_i + \boldsymbol{U}\boldsymbol{x}_{i1})$ and $(\boldsymbol{m} + \boldsymbol{d}_i + \boldsymbol{U}\boldsymbol{x}_{i2})$, which are clearly closer to the ML points, $ML(\boldsymbol{O}_{i1})$ and $ML(\boldsymbol{O}_{i2})$.

### 3.5. *Classification*

Classification for ISV and JFA still uses an LLR score similar to (12). The key difference is that we know there is unwanted session variation in the test images that we want to compensate for. In the previous two sections we discussed how this unwanted session variation was excluded during enrolment of the client models for ISV and JFA. In this section we discuss how to incorporate an estimate of the session variation in a test image during LLR scoring.

A method to compensate for the effects of session variation in a test image $\boldsymbol{O}_t$ (or set of observations) was proposed in [5]. Given a model for the $i^{th}$ client without session variability effects ($\boldsymbol{s}_i^{ISV}$ for ISV and $\boldsymbol{s}_i^{JFA}$ for JFA) we estimate the latent session variable $\boldsymbol{x}_{i,t}$ for image $\boldsymbol{O}_t$. Using this estimated latent session variable we apply the corresponding offset to the $i^{th}$ client model (thus, $\boldsymbol{s}_i^{ISV} + \boldsymbol{U}\boldsymbol{x}_{i,t}$ for ISV and $\boldsymbol{s}_i^{JFA} + \boldsymbol{U}\boldsymbol{x}_{i,t}$ for JFA). This explicitly compensates for the estimated noise in $\boldsymbol{O}_t$ because we then evaluate the likelihood that the observed image was produced by the claimed identity, $i$, *in the estimated noise conditions*. Extending this to the case of the UBM results in the new LLR,

$$h\left(\boldsymbol{O}_t, \boldsymbol{s}_i^*\right) = \sum_{k=1}^{K} \left(\ln(p(\boldsymbol{o}_t^k \mid \boldsymbol{s}_i^* + \boldsymbol{U}\boldsymbol{x}_{i,t})) - \ln(p(\boldsymbol{o}_t^k \mid \boldsymbol{m} + \boldsymbol{U}\boldsymbol{x}_{UBM,t}))\right), \quad (40)$$

where we have used $\boldsymbol{s}_i^*$ to indicate either $\boldsymbol{s}_i^{ISV}$ or $\boldsymbol{s}_i^{JFA}$.

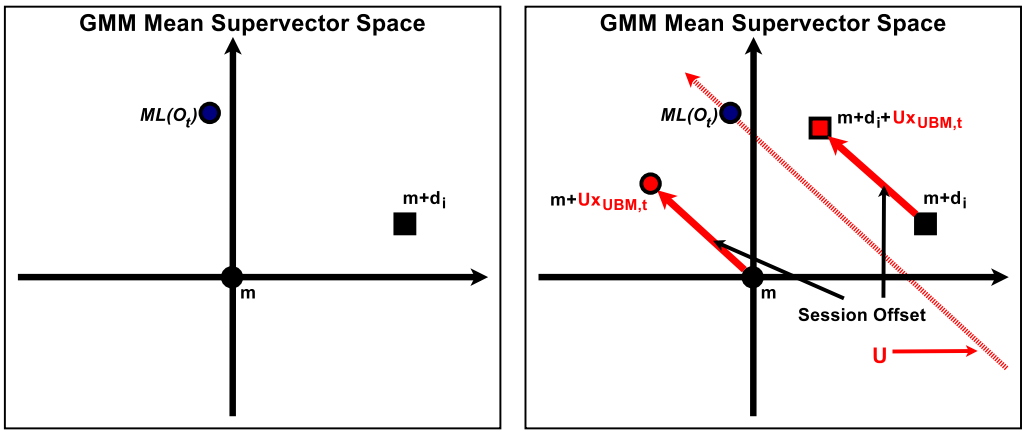
It was shown in [16] that a series of simplifications can be applied including the important assumption that the latent session variable for image $\boldsymbol{O}_t$ for each client can be approximated using the UBM. By doing this, for each image, we only need to compute one latent session offset $\boldsymbol{U}\boldsymbol{x}_{UBM,t}$. This assumption is referred to as the LPT assumption and changes the LLR (40) to be,

$$h\left(\boldsymbol{O}_t, \boldsymbol{s}_i^*\right) = \sum_{k=1}^{K} \left(\ln(p(\boldsymbol{o}_t^k \mid \boldsymbol{s}_i^* + \boldsymbol{U}\boldsymbol{x}_{UBM,t})) - \ln(p(\boldsymbol{o}_t^k \mid \boldsymbol{m} + \boldsymbol{U}\boldsymbol{x}_{UBM,t}))\right). \tag{41}$$

Continuing with the linear scoring simplifications the final LLR is approximated by,

$$h_{\text{linear}}\left(\boldsymbol{O}_t, \boldsymbol{s}_i^*\right) = \left(\boldsymbol{s}_i^* - \boldsymbol{m}\right)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{f}_{t|\boldsymbol{m}}. \tag{42}$$

An illustration of how this form of scoring works is given in Figure 3.



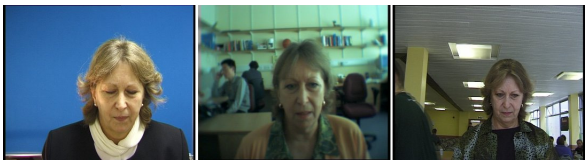

(a) Without a session variability model. (b) With a session variability model.

Figure 3: This is an illustration of how scoring is performed (a) *without* a session variability model is performed and (b) *with* a session variability model. In both cases we assume that we are given the world model ($\boldsymbol{m}$) and a model for the claimed client ($\boldsymbol{m} + \boldsymbol{d}_i$) and that we want to calculate a score (LLR) for the image $\boldsymbol{O}_t$, represented here by its maximum likelihood (ML) point given by (39). In (b), we show that when we include a session variability subspace, $\boldsymbol{U}$, we estimate the noise present in $\boldsymbol{O}_t$ resulting in a session offset $\boldsymbol{U}\boldsymbol{x}_{UBM,t}$. We then compensate for the estimated noise in the image by scoring against the compensated world model ($\boldsymbol{m} + \boldsymbol{U}\boldsymbol{x}_{UBM,t}$) and client model ($\boldsymbol{m} + \boldsymbol{d}_i + \boldsymbol{U}\boldsymbol{x}_{UBM,t}$).

## 4. Experimental protocols

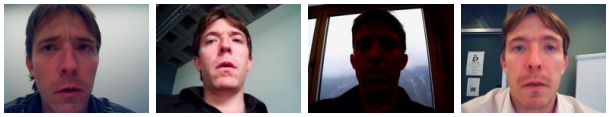
For this article we perform experiments on several publicly-available face authentication databases. To properly evaluate ISV and JFA, we chose to use images taken in challenging conditions causing substantial within-class variation. Furthermore, we chose to restrict ourselves to publicly-available databases with sepa-

rate training, development and evaluation sets to allow for unbiased evaluation. As mentioned in our prior work [13] some popular databases such as FRGC [20] and LFW [24] were thus not applicable as they do not include separate development and evaluation sets[3]. We therefore chose to evaluate the ISV and JFA techniques on the challenging BANCA, SCface, MOBIO and Multi-PIE databases. Example images from these databases are provided in Figure 4. We describe a brief description of these four database and the challenges they present below.



(a) BANCA database.

(b) SCface database



(c) MOBIO database.

(d) Multi-PIE database.

Figure 4: Example images showing a wide range of within-client variation (session variability) for (a) BANCA, (b) SCface, (c) MOBIO and (d) Multi-PIE.

[3]In the FRGC database, 153 clients occur in both the training set as well as the evaluation 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 evaluation set (*View 2*).

**BANCA [25]** is an access control database that has three environmental conditions, *controlled*, *degraded* and *adverse*, and also includes acquisition device (camera) variation. We use the Pooled test (P) on the English subset for both manual (*Manual*) and automatic (*Auto.*) eye annotations based on the detector of [26]. Using the automatic annotations from a face detector allowed us to test the impact of real-world face misalignment.

**SCface [27]** is a surveillance database that has four sets of images: *high resolution*, *close*, *medium* and *far*. We use the Combined (*Comb.*) protocol defined in our prior work [13][4]. This allows us to present the performance across a range of surveillance tasks.

**MOBIO [3]** is a database that was captured almost exclusively on mobile phones. The clients held the mobile device which they talked to by answering a series of prompted questions. As such the database includes pose and significant illumination variation. We use the well defined protocols that come with this database to perform separate *Male* and *Female* trials on the MOBIO Still-Image database[4]; we do gender independent training.

**Multi-PIE [4]** is a large database consisting of 337 identities with varying facial expression and significant pose and illumination variation. We used this large database to examine the effectiveness of JFA when there was a large number of clients available in the training set. To the best of our knowledge, there is no face authentication protocol nor manual annotations available for this database. For this reason we manually annotated the illumination varying part of this database, using frontal images from camera

[4]The protocol and manual annotations are available from http://www.idiap.ch/resource/biometric

05_1, and defined a protocol. The protocol is referred to as the Multi-PIE Face Verification Protocol – Unmatched Illumination[4].

To assess face authentication accuracy we use a consistent procedure for all of the databases. For each database three independent sets are defined: (i) a training set, (ii) a development set, and (iii) an evaluation set. The three sets were made independent by ensuring that no client in one set occurred in any other set. The training set was used to learn parameters for models such as training the UBM GMM as well as the subspaces for ISV and JFA. The development set was used to derive the optimal hyper-parameters for these models such as the number of components, number of dimensions as well as deriving a global decision threshold to minimise the equal error rate (EER), which is the operating point of interest in this work. This threshold was then applied to an evaluation 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 hyper-parameters, were tuned prior to seeing the evaluation set. In Table 1 we summarise the contents of each database.

## 5. Results

Prior work in [13] showed that ISV and JFA could be successfully applied to face authentication. It was found that JFA was not as effective as ISV for face authentication, and it was conjectured that one possible reason for this was that there were only a small number of clients in the training set causing the identity subspace $\boldsymbol{V}$ to not be adequately trained. We examine this hypothesis by performing experiments on the much larger Multi-PIE database, which contains 208 clients for training. We also note that in the prior work [13] the training of the ISV subspace $\boldsymbol{U}$ was performed without estimating the latent variable $\boldsymbol{z}_i$, which was

| | Training Set | Development Set | | Evaluation Set | |
|---|---|---|---|---|---|
| | | ***Enrolment*** | ***Testing*** | ***Enrolment*** | ***Testing*** |
| | clients (images) | clients (images/client) | true trials, false trials | clients (images/client) | true trials, false trials |
| Multi-PIE | 208 clients (9,785) | 64 (1 image) | 4,864 true, 306,432 false | 65 (1 image) | 4,940 true, 316,160 false |
| MOBIO *(Male)* | 50 clients (9,579) | 24 (5 images) | 2,520 true, 57,960 false | 38 (5 images) | 3,990 true, 147,630 false |
| MOBIO *(Female)* | 50 clients (9,579) | 18 (5 images) | 1,890 true, 32,130 false | 20 (5 images) | 2,100 true, 39,900 false |
| BANCA | 30 clients (300) | 26 (5 images) | 1,170 true, 1,560 false | 26 (5 images) | 1,170 true, 1,560 false |
| SCface | 43 clients (688) | 44 (1 image) | 660 true, 28,380 false | 43 (1 image) | 645 true, 27,090 false |

Table 1: Above is a summary of the four databases used in these experiments. For training, we have included the number of clients for training (clients) and the overall number of images for training (images). For Development and Evaluation we have included the number of clients for enrolment (clients), the number of enrolment images per client (images/client) and the overall number of true and false trials (true trials, false trials).

an unintentional error. While this error does not change the overall conclusions of the prior work [13], for the sake of consistency and transparency we have chosen to re-run all of the ISV and JFA experiments for this article. In addition to this we compare the performance of all of our systems against the related state-of-the-art MRH approach [12]. The systems described in this paper make use of and have been implemented in *bob* [28] [5].

When performing the experiments, we first derive a baseline GMM system for each database before training the ISV and JFA models. The hyper-parameters

[5] This is a signal-processing and machine learning toolbox that is freely available for research purposes http://www.idiap.ch/software/bob/.

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 and a relevance factor of $\tau = 4$ was used for client model adaptation. The subspaces were trained using 10 EM iterations and the 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 [5]. Manual face localisation was used unless otherwise noted.

### *5.1. GMM and MRH baseline systems*

The GMM baseline system that we use was optimised in a similar manner as outlined in [13]. The size of the blocks used during feature extraction was tuned on the development set. The number of 2D-DCT coefficients for a given block size, $O$, was initially tuned on BANCA and we re-used the results of [13] which found that the optimal block sizes were $12 \times 12$ pixels with $M = 44$ and $20 \times 20$ pixels with $M = 65$ for the BANCA and SCface databases, respectively. For MOBIO and Multi-PIE a block size of $12 \times 12$ was chosen with $M = 44$.

In addition to optimising the block size we also examined the issue of using ZT-norm score normalisation which is not yet common place in face authentication but in [13, 29] it was found to be advantageous. We conducted a set of experiments on all four databases to confirm this. The ZT-norm set for BANCA and Multi-PIE came from the evaluation set when optimising for the development set and vice versa when optimising for the evaluation set. For the SCface and MOBIO databases, two-thirds of the identities in the training set were used for deriving the UBM and the remaining one-third were used for ZT-norm. The summary of results for ZT-norm score normalisation for the baseline GMM systems, provided in Table 2, confirm that ZT-norm generally provides a significant improvement in

performance with an average relative improvement of $30\%$. However, ZT-norm did obtain mixed results for MOBIO where it helped on the evaluation set for the *Male* trials but did not help for the *Female* trials. Given the overall benefit of ZT-norm we decided to use it with the GMM baseline system and all ISV and JFA systems for the remainder of the experiments. This GMM baseline system with tuned block size and ZT-norm score normalisation is kept the same for the remainder of the experiments.

Using the optimal GMM baseline system we then evaluated the performance of MRH. Using the same features as the GMM baseline, it was found for MRH that ZT-norm score normalisation provided a consistent performance improvement, with an average relative performance improvement of $31\%$. However, it was also found that MRH was outperformed by the GMM baseline for all of the databases with the GMM baseline providing an average relative improvement of $30\%$. We attribute this performance difference to the fact that MRH only uses the zeroth order statistics of the GMM UBM which are then compared using an L1 distance, by comparison the GMM baseline incorporates both the zeroth and first order statistics of the GMM UBM which are then compared using a Mahalanobis distance (13), see Section 2.4. We also note that the MRH system was never compared to a GMM baseline in [12] and that we believe that the performance of the GMM baseline could also be improved if we consider a multi-region approach, however, this is beyond the scope of this paper.

### *5.2. Inter-session variability modelling*

To evaluate the accuracy of ISV we first tuned the dimensionality, $n_x$, of the subspace $\boldsymbol{U}$ on the development set of each database using $n_x = [5, 10, 20, 40, 80, 160, 320]$. We again note that the previous work [13] contained

| | MRH [12] | | | | GMM | | | |
|---|---|---|---|---|---|---|---|---|
| | Without ZT-norm | | With ZT-norm | | Without ZT-norm | | With ZT-norm | |
| | Devel. EER % | Eval. HTER % | Devel. EER % | Eval. HTER % | Devel. EER % | Eval. HTER % | Devel. EER % | Eval. HTER % |
| Multi-PIE | 12.2 | 13.8 | 4.8 | 6.2 | 7.6 | 6.7 | 3.0 | **3.2** |
| MOBIO *(Male)* | 13.5 | 17.4 | 13.6 | 13.0 | 8.9 | **11.9** | 9.2 | 10.5 |
| MOBIO *(Female)* | 12.9 | 22.6 | 14.5 | 21.9 | 10.3 | **18.2** | 10.7 | 20.4 |
| BANCA *(Manual)* | 14.3 | 13.8 | 9.3 | 8.4 | 11.0 | 11.1 | 7.8 | **6.1** |
| BANCA *(Auto.)* | 17.4 | 15.4 | 11.2 | 10.1 | 13.6 | 12.5 | 9.2 | **6.7** |
| SCface *(Comb.)* | 42.6 | 42.5 | 28.3 | 30.3 | 23.9 | 25.1 | 16.7 | **16.4** |

Table 2: We present the results for the baseline GMM and MRH systems with and without ZT-norm. In bold, we have highlighted the result on the evaluation set (Eval.) for the best system on the development set (Devel.).

an error in the training of the subspace $\boldsymbol{U}$, for ISV only, and so the new results reported here supersede the prior results.

It was found that ISV provides a consistent improvement in performance compared to the GMM baseline, as shown in Table 3. This improvement occurs for both the development and evaluation sets and on average, across all of the databases, provides a relative improvement of $26\%$ with the minimum relative improvement being $11\%$ on BANCA (Manual) and the maximum relative improvement being $40\%$ on MOBIO (Female), reducing the HTER from $20.4\%$ to $12.2\%$. This has shown that ISV provides significant performance gains in the presence of a number of different forms of session variation which are present in the databases that we have evaluated on. In particular, it provides state-of-the-art performance on the SCface surveillance database. When evaluating specifically for illumination variation, on the Multi-PIE database, it provided an impressive performance gain of $38\%$. Also, for the case of real-world misalignment, that can be caused when using the results from an automatic face detector, it was shown

on the BANCA database that the relative reduction in performance from BANCA (Manual) to BANCA (Auto.) is only $6\%$ for ISV whereas the GMM and MRH systems have a relative reduction in performance of $10\%$ and $20\%$ respectively [6].

### *5.3. Joint factor analysis*

In [13], it was hypothesised that the lack of identity diversity could have led to JFA providing inferior performance. Therefore, for these extended experiments we examined the impact of using the Multi-PIE database because it has 208 identities in the training set. If JFA was really limited by the number of different training identities the large Multi-PIE training set should provide JFA with a significant boost in performance with respect to the databases with smaller training sets. When running these extended experiments we tuned the dimensionality of the $\boldsymbol{U}$ subspace using $n_x = [5, 10, 20, 40, 80, 160, 320]$ and the $\boldsymbol{V}$ subspace using $n_y = [5, 10, 20, 40, 80, 160]$ (up to a maximum value of the number of distinct identities in the training set minus one).

The results in Table 3 show that JFA can sometimes provide improved performance over the baseline GMM system. JFA provides marginally better performance than ISV for MOBIO (male) on the evaluation set only and for SCface for the development set only, however, in every other case ISV outperforms JFA. On average, for all of the databases, JFA provides a relative improvement of $12\%$ over the baseline GMM system which is half the average relative improvement of ISV. More critically, JFA does not perform well on the large Multi-PIE database. On Multi-PIE, JFA performs worse than ISV and only marginally better than the GMM Baseline. When compared to the GMM Baseline, JFA obtains a relative improvement of $3\%$ compared to ISV which obtains a relative improvement of

[6] The high degradation in performance for MRH is likely due to the use of a fixed $3 \times 3$ grid.

38%. Given this result we chose to run supplementary experiments on Multi-PIE to further examine the effect of varying the number of identities in the training set for JFA. The aim of these experiments was to discover if JFA was potentially being limited by not having sufficient diversity (number of identities) in the training data set. To do this we trained JFA systems on a reduced number of identities consisting of 41, 81, 121 and 161 identities by randomly selecting them from the full training set of 208. Figure 5 shows that JFA consistently improves when we add more identities to the training set and suggests that an even larger training data set could improve the accuracy of JFA further. This also highlights a significant advantage of ISV, that is, the ability to provide good improvements over the baseline with a modest, and perhaps more realistic, amount of training data.

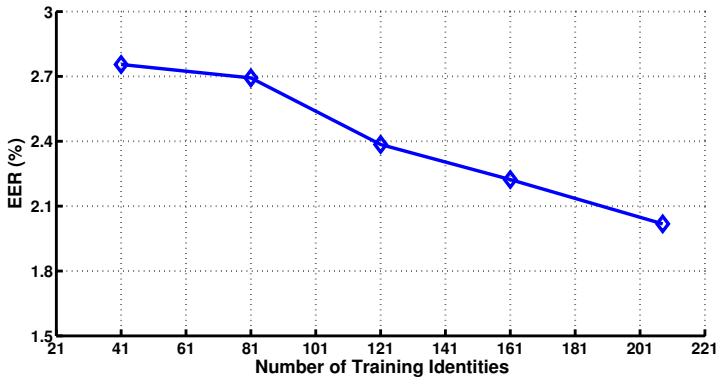



Figure 5: In the above plot we provide the effect, on the development set of Multi-PIE, of varying the number of training identities to train JFA. It can be seen that the EER consistently decreases as we add more identities to the training set.

## 6. Analysis of variation in subspaces

To better understand the performance of ISV and JFA we designed another set of experiments to analyse the amount of between-class and within-class information that these two models were capturing.

| | MRH [12] | | GMM | | ISV | | JFA | |
|---|---|---|---|---|---|---|---|---|
| | Dev. EER % | Eval. HTER % | Dev. EER % | Eval. HTER % | Dev. EER % | Eval. HTER % | Dev. EER % | Eval. HTER % |
| Multi-PIE | 4.8 | 6.2 | 3.0 | 3.2 | 1.4 | **2.0** | 2.0 | 3.1 |
| MOBIO *(Male)* | 13.6 | 13.0 | 9.2 | 10.5 | 3.6 | **7.5** | 4.1 | 7.4 |
| MOBIO *(Female)* | 14.5 | 21.9 | 10.7 | 20.4 | 6.7 | **12.2** | 7.8 | 15.5 |
| BANCA *(Manual)* | 9.3 | 8.4 | 7.8 | 6.1 | 6.6 | **5.4** | 7.7 | 6.1 |
| BANCA *(Auto.)* | 11.2 | 10.1 | 9.2 | 6.7 | 8.0 | **5.7** | 9.2 | 6.7 |
| SCface *(Comb.)* | 28.3 | 30.3 | 16.7 | 16.4 | 12.8 | 13.0 | 12.0 | **13.5** |

Table 3: We present the results for the multi-region histogram approach [12], Baseline GMM, ISV and JFA systems on the Multi-PIE, MOBIO (Male and Female sets), BANCA (manual and automatic annotations) and SCface databases. In bold we have highlighted the result on the evaluation set (Eval.) for the best system chosen from the development set (Dev.), in all cases we use ZT-norm.

ISV and JFA both aim to have reliable estimates of the real client identity with limited data. They both use a session-specific subspace $\boldsymbol{U}$ to capture and suppress within-class (WC) variation while still retaining as much between-class (BC) variation as possible, in order to discriminate between different people. We use this as the basis for analysing the performance of ISV and JFA.

To analyse the performance of ISV and JFA we evaluate the amount of WC and BC variation that is captured in the client-dependent part of the model $\boldsymbol{d}_i$; for ISV this is $\boldsymbol{D}\boldsymbol{z}_i$ (19) and for JFA this is $\boldsymbol{V}\boldsymbol{y}_i + \hat{\boldsymbol{D}}\boldsymbol{z}_i$ (22). To quantify how reliably these terms are estimated by the model we propose the following procedure. For a set of images disjoint from the training set (we use the evaluation set enrolment and test images) we do the following.

1. Find the MAP estimates of $\boldsymbol{d}_i$ and $\boldsymbol{u}_{i,j}$, but without the constraint that $\boldsymbol{d}_i$ is constant for images of the same person. To make this distinction clear we refer to the estimates of $\boldsymbol{d}_i$ made in this way as $\hat{\boldsymbol{d}}_{i,j}$.
2. Calculate the proportion of WC variation in $\hat{\boldsymbol{d}}_{i,j}$ with respect to the WC

variation in $\hat{\boldsymbol{\mu}}_{i,j} = \hat{\boldsymbol{d}}_{i,j} + \boldsymbol{u}_{i,j}$. This defines a WC ratio,

$$\sigma^w = \frac{tr\left(S_w(\hat{\boldsymbol{d}}_{i,j})\right)}{tr\left(S_w(\hat{\boldsymbol{\mu}}_{i,j})\right)}. \tag{43}$$

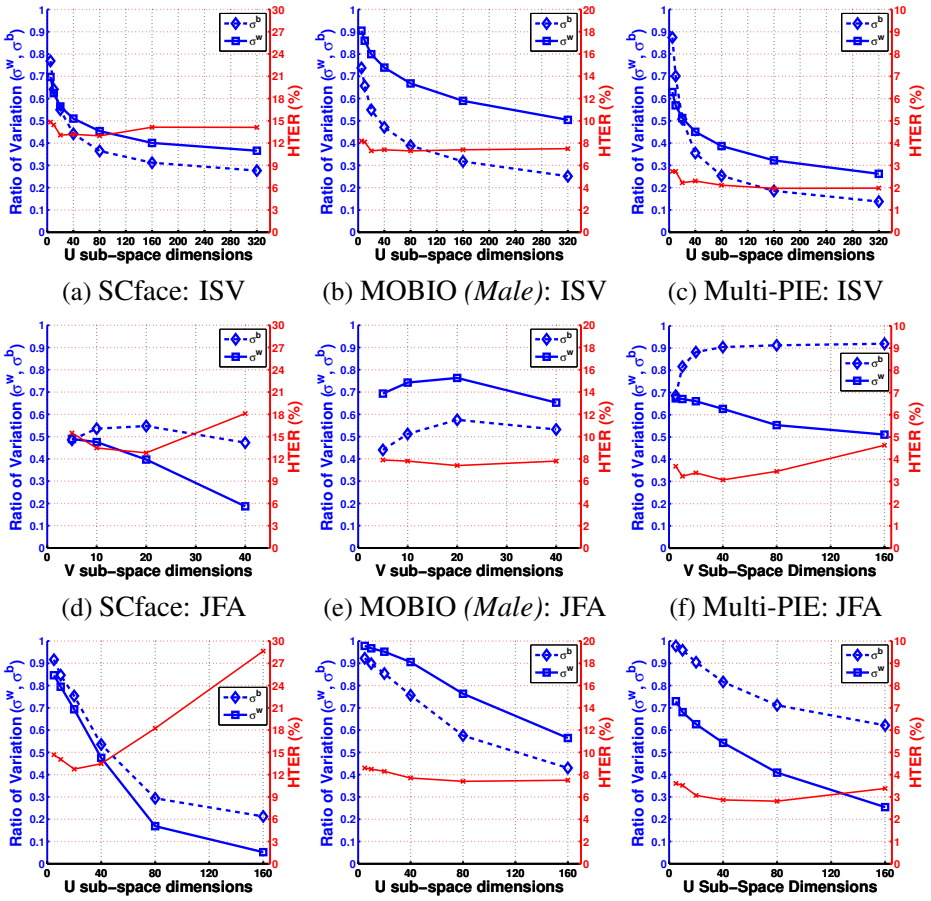
The super-vectors $\hat{\boldsymbol{\mu}}_{i,j}$ are composed of both the client-dependent and session-dependent parts of the model, thus the WC variation in $\hat{\boldsymbol{\mu}}_{i,j}$ represents the overall WC variation observed in the data set. If the subspace $\boldsymbol{U}$ (plus $\boldsymbol{V}$ and $\hat{\boldsymbol{D}}$ for JFA) is well-estimated, we hypothesise that $\hat{\boldsymbol{d}}_{i,j}$ should be close to constant for images of the same person and thus, the WC variation in $\hat{\boldsymbol{d}}_{i,j}$ should be much less than that in $\hat{\boldsymbol{\mu}}_{i,j}$. Consequently, a well-estimated model should give rise to a small WC ratio $\sigma^w$ and thus be expected to improve accuracy.

3. Calculate the proportion of BC variation in $\hat{\boldsymbol{d}}_{i,j}$ with respect to the overall variation in $\hat{\boldsymbol{\mu}}_{i,j}$. Similar to the WC ratio, the BC ratio is

$$\sigma^b = \frac{tr\left(S_b(\hat{\boldsymbol{d}}_{i,j})\right)}{tr\left(S_b(\hat{\boldsymbol{\mu}}_{i,j})\right)}. \tag{44}$$

Converse to the WC ratio (43), the BC ratio $\sigma^b$ should be large for systems that are well trained and thus be expected to improve accuracy.

In Figures 6 (a)-(c), we present the results of the analysis of ISV for three databases. It can be seen that $\sigma^w$ decreases as we increase the size of the session subspace $\boldsymbol{U}$. This suggests that the session variability term $\boldsymbol{u}_{i,j}$ removes WC variation from the client models and thus explains the improved HTER with respect to the baseline system. At the same time it also reduces $\sigma^b$ which is undesirable, however, this does not come at the cost of reduced accuracy as for all of the graphs the HTER is not overly sensitive to the size of the $\boldsymbol{U}$ subspace.




(a) SCface: ISV (b) MOBIO *(Male)*: ISV (c) Multi-PIE: ISV

(d) SCface: JFA (e) MOBIO *(Male)*: JFA (f) Multi-PIE: JFA

(g) SCface: JFA (h) MOBIO *(Male)*: JFA (i) Multi-PIE: JFA

Figure 6: We present the analysis plots of $\sigma^w$ and $\sigma^b$, in solid and dashed blue lines respectively, along with the associated HTER (%), in red, for SCface, MOBIO (*Male*) and Multi-PIE. From top to bottom we have the result for: (a)-(c) ISV by varying the dimensionality of $\boldsymbol{U}$, (d)-(f) JFA using the optimal size for $\boldsymbol{U}$ and varying the dimensionality of $\boldsymbol{V}$, and (g)-(i) JFA using the optimal size for $\boldsymbol{V}$ and varying the dimensionality of $\boldsymbol{U}$ (g)-(i).

For JFA we analyse the effect of varying the identity subspace $\boldsymbol{V}$ in Figures 6 (d)-(f). We retain the optimal session subspace $\boldsymbol{U}$ from the development set. It can be seen that up to a certain point, as we increase the size of the identity subspace we obtain an improved HTER and the BC ratio $\sigma^b$ increases. This improvement in HTER stops for SCface and MOBIO *Male* when $n_y = 20$ and for Multi-PIE when

$n_y = 40$. After these points the HTER worsens (increases) and the BC ratio $\sigma^b$ decreases or in the case of Multi-PIE plateaus. This demonstrates that although JFA appears to be working for small subspaces, it does not appear to be more effective when we use large identity subspaces. Unfortunately, using these small identity subspaces does not result in improved performance when compared to the simpler ISV model.

To extend our analysis of JFA we also examined the effect of varying the size of the session subspace $\boldsymbol{U}$, as shown in Figures (g)-(i). A similar trend to ISV (Figures (a)-(c)) can be seen for the WC ratio $\sigma^w$ and the BC ratio $\sigma^b$. As we increase the size of $\boldsymbol{U}$ we decrease the WC ratio $\sigma^w$, but we also decrease the BC ratio $\sigma^b$. It can be seen that for MOBIO and Multi-PIE, Figures (h) and (i), increasing the session subspace size initially improves the HTER but this quickly plateaus, and overall it is not too sensitive to this parameter. However, for SCface, which has a limited training data set, increasing the session subspace quickly results in a significant degradation of the HTER. This suggests that when limited training data is used for JFA it is more sensitive to the size of $\boldsymbol{U}$.

Our analysis of $\sigma^w$ and $\sigma^b$ has highlighted a limitation common to both ISV and JFA for face authentication. This limitation stems from the fact that the session subspace $\boldsymbol{U}$ suppresses a significant amount of BC variation $\sigma^b$ along with the WC variation $\sigma^w$. A similar problem was recently found for speaker authentication and was one of the motivations for an alternative approach to factor analysis of GMM mean super-vectors, referred to as total variability modelling [30]. In this alternative approach, the GMM mean super-vectors were represented by low-dimensional feature vectors which were able to be more effectively modelled using session variability modelling techniques. A similar approach may be worth

investigating for face authentication.

## 7. Conclusions

This work has shown that session variability modelling can lead to significant improvements in face authentication accuracy. We have shown that inter-session variability modelling (ISV) consistently improves the relative performance by on average $26\%$, when compared to the GMM Baseline system. By comparison JFA provides an average relative improvement of $12\%$.

A comparison with the state-of-the-art MRH system [12] has also been provided. Experimentally it was found that all of the systems, GMM, ISV and JFA, consistently outperformed the MRH approach, on average the GMM baseline system provided a relative performance improvement of $30\%$ compared to the MRH system. We provide an insight as to why the GMM systems outperform MRH by noting that the MRH system relies on the zeroth order statistics of the GMM UBM whereas the GMM systems incorporate both the zeroth and first order statistics.

We have also provided the first self-contained detailed description of the ISV and JFA algorithms for face authentication and provided all of the equations necessary to implement these two techniques. We have reported results on several publicly-available databases and provided a novel face authentication protocol for the large new Multi-PIE database and an expanded set of results since the initial work of [13].

In addition to this, to better understand both ISV and JFA we have analysed the WC and BC variation that these two models capture. In analysing the subspaces for ISV and JFA we have highlighted a common limitation. Results suggest that the session subspace $\boldsymbol{U}$ inadvertently ends up modelling and suppressing a significant amount of BC variation along with the WC variation. To overcome this

problem we suggest that, following recent advances in speaker authentication, future work should examine total variability modelling [30] so that the GMM mean super-vectors are represented by low-dimensional i-vectors which can then be effectively modelled using session variability modelling techniques.

## 8. Acknowledgements

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) and from NICTA. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program. Portions of this research 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.

## References

[1] E. Rúa, J. Castro, and C. Mateo, "Quality-based Score Normalization for Audiovisual Person Authentication," in *Image Analysis and Recognition*, ser. Lecture Notes in Computer Science, 2008, vol. 5112, pp. 1003–1012. 2

[2] T. Ahonen and M. Pietikäinen, "Pixelwise Local Binary Pattern Models of Faces using Kernel Density Estimation," in *Advances in Biometrics*, ser. Lecture Notes in Computer Science, 2009, vol. 5558, pp. 52–61. 2

[3] C. McCool *et al.*, "Bi-Modal Person Recognition on a Mobile Phone: using mobile phone data," *IEEE International Conference on Multimedia and Expo Workshop on Hot Topics in Mobile Multimedia*, 2012. 2, 24

[4] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image and Vision Computing*, vol. 28, pp. 807–813, 2010. 2, 24

[5] R. Vogt and S. Sridharan, "Explicit Modelling of Session Variability for Speaker Verification," *Computer Speech & Language*, vol. 22, no. 1, pp. 17–38, 2008. 2, 8, 11, 12, 13, 15, 16, 17, 18, 21, 27

[6] 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*, vol. 15, no. 4, pp. 1435–1447, 2007. 2, 11, 14

[7] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000. 3, 6, 7, 8, 9

[8] C. Sanderson and K. Paliwal, "Fast Features for Face Authentication under Illumination Direction Changes," *Pattern Recognition Letters*, vol. 24, pp. 2409–2419, 2003. 3, 5

[9] 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*, vol. 2, 2004, pp. 855–861. 3, 5, 7

[10] F. Cardinaux, C. Sanderson, and S. Bengio, "User Authentication via Adapted Statistical Models of Face Images," *IEEE Transactions on Signal Processing*, vol. 54, no. 1, pp. 361–373, 2006. 3, 5, 7

[11] C. McCool and S. Marcel, "Parts-based Face Verification using Local Frequency Bands," in *Proceedings of the Third International Conference on Advances in Biometrics*, 2009. 3

[12] C. Sanderson and B. C. Lovell, "Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference," in *Proceedings of the Third International Conference on Advances in Biometrics*, 2009. 3, 4, 5, 10, 26, 28, 29, 32, 36

[13] R. Wallace, M. McLaren, C. McCool, and S. Marcel, "Inter-session Variability Modelling and Joint Factor Analysis for Face Authentication," in *Proceedings of the International Joint Conference on Biometrics*, 2011, pp. 1–8. 3, 4, 23, 24, 25, 26, 27, 28, 30, 36

[14] 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*, ser. Lecture Notes in Computer Science, 2003, vol. 2688, pp. 1058–1059. 5

[15] X. Tan and B. Triggs, "Enhanced Local Texture Feature Sets for Face Recognition under Difficult Lighting Conditions," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1635–1650, 2010. 5

[16] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, "Comparison of Scoring Methods used in Speaker Recognition with Joint Factor Analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 4057–4060. 9, 21

[17] O. Thyes, R. Kuhn, P. Nguyen, and J. Junqua, "Speaker Identification and Verification using Eigenvoices," in *Proceedings of the International Conference on Spoken Language Processing*, vol. 2, 2000, pp. 242–245. 14

[18] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A Study of Inter-Speaker Variability in Speaker Verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, pp. 980–988, 2008. 16, 19

[19] L. Burget, M. Fapšo, V. Hubeika, O. Glembek, M. Karafiát, M. Kockmann, P. Matějka, P. Schwarz, and J. Černocký, "BUT System Description: NIST SRE 2008," in *Proceedings of the 2008 NIST Speaker Recognition Evaluation Workshop*, 2008, pp. 1–4. 19

[20] 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 *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 947–954. 23

[21] P. Phillips, J. Beveridge, B. Draper, G. Givens, A. O'Toole, D. Bolme, J. Dunlop, Y Lui, H. Sahibzada, and S. Weimer, "An Introduction to the Good, the Bad, & the Ugly Face Recognition Challenge Problem," in *Proceedings of the IEEE International Conference on Face and Gesture*, 2011, pp. 346–353. 3

[22] J. Yang, A. Frangi, J. Yang, D. Zhang, and Z. Jin, "KPCA Plus LDA: A Complete Kernel Fisher Discriminant Framework for Feature Extraction and

Recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, vol. 27, issue. 2, pp. 230–244. 3

[23] M. Günther, R. Wallace, and S. Marcel, "An Open Source Framework for Standardized Comparisons of Face Recognition Algorithms," in *Proceedings of the Second IEEE International Workshop on Benchmarking Facial Image Analysis Technologies, in conjunction with ECCV 2012*, To appear. 3

[24] 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." University of Massachusetts, Amherst, Tech. Rep. 07-49, 2007. 23

[25] E. Bailly-Baillière *et al.*, "The BANCA Database and Evaluation Protocol," in *Audio- and Video-Based Biometric Person Authentication*, ser. Lecture Notes in Computer Science, 2003, vol. 2688, pp. 1057–1071. 24

[26] Y. Rodriguez, "Face Detection and Verification using Local Binary Patterns," Ph.D. dissertation, Idiap Research Institute and École Polytechnique Fédérale de Lausanne, 2006. 24

[27] M. Grgic, K. Delac, and S. Grgic, "SCface-Surveillance Cameras Face Database," *Multimedia tools and applications*, vol. 51, pp. 863–879, 2011. 24

[28] A. Anjos, L. El Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel, "Bob: a free signal processing and machine learning toolbox for researchers," *Proceedings of the ACM Multimedia Conference*, 2012. 26

[29] R. Wallace, M. McLaren, C. McCool, and S. Marcel, "Cross-pollination of Normalisation Techniques from Speaker to Face Authentication using Gaussian Mixture Models," *IEEE Transactions on Information Forensics and Security*, vol. 7, pp. 553–562, 2012. 27

[30] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 788–798, 2011. 35, 37