Pattern Recognition 46 (2013) 3341–3357

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A user-specific and selective multimodal biometric fusion strategy by ranking subjects

Norman Poh [a,*], Arun Ross [b], Weifeng Lee [c], Josef Kittler [d]

[a] Department of Computing, University of Surrey GU2 7XH, UK
[b] West Virginia University, Morgantown, WV 26506, USA
[c] Shenzhen Key Lab. of Information Sci&Tech/Shenzhen Engineering Lab. of IS&DRM, Department of Electronic Engineering/Graduate School at Shenzhen, Tsinghua University, China
[d] Centre for Vision, Speech and Signal Processing (CVSSP), FEPS, University of Surrey GU2 7XH, UK

ARTICLE INFO

Article history:
Received 5 October 2012
Received in revised form
28 February 2013
Accepted 22 March 2013
Available online 6 April 2013

Keywords:
Biometrics
Multibiometrics
Fusion
Doddington's Zoo
User-specific fusion
Client specific fusion

ABSTRACT

The recognition performance of a biometric system varies significantly from one enrolled user to another. As a result, there is a need to tailor the system to each user. This study investigates a relatively new fusion strategy that is both *user-specific* and *selective*. By user-specific, we understand that each user in a biometric system has a different set of fusion parameters that have been tuned specifically to a given enrolled user. By selective, we mean that only a subset of modalities may be chosen for fusion. The rationale for this is that if one biometric modality is sufficiently good to recognize a user, fusion by multimodal biometrics would not be necessary, we advance the state of the art in user-specific and selective fusion in the following ways: (1) provide thorough analyses of (a) the effect of pre-processing the biometric output (prior to applying a user-specific score normalization procedure) in order to improve its central tendency and (b) the generalisation ability of user-specific parameters; (2) propose a criterion to rank the users based solely on a training score dataset in such a way that the obtained rank order will *maximally correlate* with the rank order that is obtained if it were to be computed on the test set; and, (3) experimentally demonstrate the performance gain of a user-specific and -selective fusion strategy across fusion data sets at different values of "pruning rate" that control the percentage of subjects for whom fusion is not required. Fifteen sets of multimodal fusion experiments carried out on the XM2VTS score-level benchmark database show that even though our proposed user-specific and -selective fusion strategy, its performance compares favorably with the conventional fusion system that considers all information.

1. Introduction

1.1. Information fusion in biometrics

Biometric authentication is the process of establishing a human identity based on a person's physical or behavioral traits. An automated biometric authentication system inputs the raw biometric data of an individual, extracts a set of features from the raw data, and compares this set against the identity models residing in the database to either verify a claimed identity or determine the individual's identity. The performance of an automated biometric authentication system is typically gauged by measuring the trade-off between the false accept rate (FAR) and the false reject rate (FRR). For a given system, it is not possible to reduce these two error rates simultaneously. A promising approach to significantly decrease these two error rates is to employ more than one biometric system with each system generating its own decision about an individual's identity (or claimed identity). Such systems, known as multibiometric systems [1], reconcile the information presented by multiple sub-systems. When N independently constructed sub-systems function together, the N output scores can be consolidated into a single output. This is the problem of *score-level fusion* which is the most popular fusion approach due to the ease of accessing scores from commercial matchers. Most multibiometric systems described in the literature employ a common fusion mechanism for all users. We refer to such a fusion mechanism as *user-independent* fusion, in order to contrast it with a user-specific fusion (to be explained later). Furthermore, these systems typically assume that all N score outputs are available, hence, requiring individual sub-systems to be operational.

1.2. The Doddington's zoo effect

An automatic biometric authentication system operates by first generating a model or template to represent each user (identity)

* Corresponding author. Tel.: +44 1483 686136.
E-mail addresses: npoh@surrey.ac.uk, normanpoh@gmail.com (N. Poh), arun.ross@mail.wvu.edu (A. Ross), li.weifeng@sz.tsinghua.edu.cn (W. Lee), j.kittler@surrey.ac.uk (J. Kittler).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.03.018

enrolled in the database. During the operational phase, the system compares the input biometric sample, also referred to as a probe, with the model in the database in order to generate a score that is eventually used to render a decision about the identity of the probe sample. Recent literature has suggested that some user models (and consequently the users they represent) are systematically better (or worse) in authentication performance than others. This effect has been referred to as “Doddington's zoo” with individual users characterized by animal labels [2] as listed below:

- *sheep*: persons who can be easily recognized;
- *goats*: persons who are particularly difficult to be recognized;
- *lambs*: persons who are easy to imitate;
- *wolves*: persons who are particularly successful at imitating others.

Goats contribute significantly to the False Reject Rate (FRR) of a system while wolves and lambs increase its False Acceptance Rate (FAR). A more recent work [3] further distinguishes four other semantic categories of users by considering both the genuine and impostor match scores for the same claimed identity simultaneously.

The literature on biometric menagerie or Doddington's zoo is related to our study here the former attempts to detect and categorize users according to their statistical behaviors at the score level whereas our study here attempts to exploit these behaviors in order to optimize the decision making process, which is the ultimate objective.

Some literature suggests that Doddington's zoo or biometric menagerie may not entirely be caused by the user himself or herself. Instead, the phenomenon is likely to be associated with the template or (statistical) reference model that represents the user. Suppose that a user has four biometric samples, T_1, T_2, T_3, and T_4. Let $M(T_i, T_j)$ represent the matching score between template T_i and query T_j. Since the matching scores $M(T_1, T_2)$ and $M(T_1, T_3)$ are generated from template T_1, they are likely to be dependent on each other. This dependency is exploited by a user-specific score normalisation or fusion in order to enhance the accept/reject decision.

This paper does not consider the case where two different templates are used to represent a user. Therefore, we do not offer an explanation as to whether or not there is a dependency (positive correlation) between $M(T_1, T_2)$ and $M(T_3, T_4)$. Answering this question would address whether or not biometric menagerie is indeed user-dependent. Although this is an important research question, we do not intend to study this scientific problem in-depth.

On the contrary, we are interested to find out the generalisation ability of a user-specific strategy when the same template remains the same over a period of time. For instance, we want to find out if the dependency between $M(T_1, T_2)$ and $M(T_1, T_3)$ still holds when T_2 and T_3 have been collected with a gap of several weeks or months apart, and the template T_1 has been kept the same throughout this period. This use scenario has practical importance because this is how nearly all biometric systems operate. This is the main research topic being pursued here.

1.3. User-specific fusion

Due to the score variation which is user-dependent, a recent trend in the literature is to design a fusion classifier that differs for each user (or client), hence, addressing the Doddington's zoo effect. Such an approach is called client-specific (or user-specific) fusion. In the literature, several user-specific fusion classifiers have been proposed based on linear classifiers [4,5], Support Vector Machines [6], Gaussian-based Bayesian classifiers [7], Multi-Layer Perceptrons [8] and discriminative classifiers based on reduced polynomial expansion [9]. These publications, however, do not associate their proposed strategy with the Doddington's zoo effect, except [10].

In [4], the authors propose a linear fusion classifier whose weights are optimized by a brute-force search and are, at the same time, constrained to be as close as possible to equal weighting. An improved scheme [5] considers the weight associated to each base classifier as inversely proportional to the “d-prime” statistic which characterizes the degree of separability between the genuine and the impostor scores for a given user.

In [6], a standard support vector machine (SVM) is trained with a user-independent set of scores as well as a user-specific one. The contribution of each set of scores is controlled by the C parameter in the SVM [11] which weights the *relative influence* of the user-independent and the user-specific data.

A similar idea using a Bayesian adaptation (instead of SVM) was reported by the same author in [7]. The architecture is similar to the Gaussian mixture model (GMM) with Maximum *A Posteriori* (MAP) adaptation, i.e., the current state-of-the-art system in speaker verification [12]. Thanks to the adaptation process, user-independent data can be exploited. However, a single Gaussian component with a diagonal covariance matrix was used, essentially realizing a Naive Bayes Gaussian classifier. This can potentially reduce its applicability to combining non-Gaussian match scores.

In [8], a multi-layer Perceptron (MLP) was used with $N+1$ inputs to combine N classifiers and the additional input is user identity normalized to sum to one. While all the available training data is used, by using a standard off-the-shelf MLP, there is no mechanism that can explicitly control the contribution of the user-independent and the user-specific information.

In [9], the authors suggested that both a user-specific fusion classifier and a user-specific threshold should be used simultaneously in order to effectively use the scarce user-specific genuine scores. However, in order to train the user-specific fusion classifier, Gaussian noise is injected to increase the training sample size.

A drawback of user-specific fusion strategy is the need for a substantial amount of training match scores. For instance, in [6], at least six (model-specific) genuine samples are needed before the user-specific procedure proposed therein can outperform the baseline system while ten samples were required in [9]. Indeed, it is difficult task to design a user-specific fusion classifier with one or two genuine training scores. Consider combining two classifiers using a Bayesian classifier built using two-dimensional Gaussian density models (one for each class). In order to obtain a full covariance, one needs at least three samples.

Rather than designing a single fusion classifier, in [13], the authors proposed to learn this classifier in two stages: firstly transform each of the sub-system output using a user-specific score normalization procedure and then combine the resulting outputs using a conventional (user-independent) fusion classifier. Note that the user-specific score normalization is not the same as the study reported in [14] because in user-specific score normalization, one uses a *different* set of normalization parameters for each user whereas in [14], a common normalization function is used for all the users. Examples of user-specific score normalization procedures are Z-norm [15,16], F-norm [17], EER-norm [18], and a likelihood-ratio based normalization [19]. These methods have been considered and compared in [13]. It was observed in [13] that the use of F-norm in the two-stage fusion process gives consistently good performance across different databases despite the lack of training samples (only 2 or 3 were used).

In [10], the properties of different animals species defined by Doddington are explicitly exploited when constructing fusion. Therefore, depending on the animal membership of a user for a particular biometric modality, the final score could depend only on one modality or the other, or both. However, this approach is not scalable to more than two modalities. Indeed, the experiments report only the fusion for two biometric modalities. For instance, if a user is a goat in two biometric modalities, the two system outputs will be fused; and, similar if a user is a lamb. However, if a user is considered a goat in the first modality but a lamb in the second modality, then the output of the first modality will be used. Since there are four possible animal species in each modality, for a fusion problem of two modalities, there will require $4\times4/2=8$ enumerations (with division by two to take into account the underlying symmetry or permutation). With 3 modalities, 16 enumerations will be needed; and with 4 modalities, 32 enumerations will be needed, and so on.

Another study derives a client-specific score normalisation scheme using the minimum of dissimilar scores resulted from exhaustive pairwise comparisons among all enrolment samples from one user [20]. This user-specific parameter is subsequently incorporated into the matching score by multiplication prior to fusion.

The brief literature review above covers only user-specific fusion algorithms. Although many novel user-independent fusion algorithms have been proposed in the literature, they are not duly covered here. By using the two-stage fusion strategy as discussed in [13], effectively any user-independent fusion algorithm can be made user-specific.

By the same token of argument, developments in user-specific score normalisation can also benefit user-specific fusion. These include group-specific score normalisation [21] and more recently, discriminative user-specific procedures using logistic regression [22].

Other schemes aim at exploiting auxiliary information through a score normalisation procedure prior to fusion. Two notable schemes are cohort-based score normalisation and quality-based score normalisation. A *cohort score normalisation* scheme attempts to blindly calibrate the distortion of matching score due to varying quality in a biometric sample via the use of a set of background of impostor (non-match) subjects who are often external to an experiment database. A prominent example of this is T-norm [16] but a recent development also includes [23]. A *quality-based score normalisation* scheme [24] attempts to model the distortion of a particular class of signal quality via the explicit use of quality measures. For example, one can objectively quantify image blur, the likelihood of having detected a face image, the clarity of ridges in fingerprints or texture in an iris image. These schemes need not be considered in isolation; they can also be combined together, e.g., [25]. Whilst the above schemes are certainly important, they do not deal specifically with the user aspect, or biometric menagerie. Therefore, they are not further discussed.

Other related literature such as user-specific feature level fusion in [26] is also an important development. However, in this paper, we will only consider user-specific fusion at the score level.

1.4. Selective or sequential fusion

Another line of research in multimodal fusion dynamically chooses only a subset of systems to be combined. In the context of multiple classifiers system, this approach is called sequential, cascaded, serial, multi-stage or cost-sensitive decision fusion [27–31]. For the purpose of this paper, we refer to all of them as selective fusion because only a subset of classifier (biometric system) outputs are used in the fusion process. A number of selective fusion strategies have been proposed for multimodal biometrics, e.g., [32–34]. All these systems have a common mechanism to determine at which stage in the sequence a decision can be made confidently. A principled approach called sequential probability ratio test was investigated by Allano and Takahashi [34,35]. However, to our knowledge, none of the sequential or selective fusion strategies studied so far is user-specific. As will become clear in this paper, doing so is extremely challenging. The solution requires intricate considerations of various aspects, such as the variability of the system performance across the subjects and the learnability of a user-ranking criterion based solely on a design data set.

1.5. Our proposal and contributions

In this paper, we propose to a framework which exploits three strategies in order to advance the state-of-the-art in *user-specific* multimodal biometric fusion. They are enumerated below.

- *User-specific normalization to facilitate multimodal fusion*: Following [13], we shall use a two-stage fusion strategy in order to obtain a user-specific fusion classifier. Such an approach has two significant advantages. Firstly, much fewer training data is needed since the normalization is a one-to-one mapping function and not multi-dimensional as is encountered in the direct approach of designing a fusion classifier, e.g., [4–6,9]. Secondly, by having a single fusion classifier, in our approach, all available (user-independent) training score data are effectively used. This is significantly different from direct approaches such as [6,7] where only some (and not all) of the user-specific data was used. Thirdly, inherent in user-specific score normalization, the parameters of different users are effectively shared by means of maximum *a posteriori* (MAP) adaptation, i.e., the normalizing parameters are adapted from a common set of parameters. As a result, the normalizing function varies slightly across users while at the same time its behavior can be globally controlled. Note that although the two-stage training approach has been examined in speaker verification [13], it has not been examined in multimodal biometric fusion as will be reported in this paper.
- *Development of a robustness criterion to rank users according to their performance*: In the second step, we introduce a criterion to rank users which will correlate very well with their rank ordering even on *unseen* data, e.g., data acquired across different sessions. This is in contrast to our prior work reported in [36] which ranks the users *a posteriori* based on observed data. Examples of criteria reported therein are Fisher-ratio [37], F-ratio (for “Fisher-like” ratio) [17], and d-prime [38]. By sorting the user templates/models, one can actually identify the ‘weak’ models. By weak models, we understand that the model consistently gives low recognition performance. Put differently, its genuine and impostor score distributions significantly overlap each other. Although there are often very few weak models, it is observed that they can disproportionately contribute to high number of recognition errors [2,39]. Weak users are made up of two species: lambs who will contribute to high false acceptance errors, and goats who will contribute to high false rejection errors. By employing discriminatory criteria, our approach diverges from [2] because lambs and goats are not distinguished. This is not a weakness because corrective actions to be taken are often similar.[1] To rank the user models, it turns out that an effective solution is to use the parameters in user-specific score normalization since these parameters can effectively gauge the user-specific score variability information.

[1] Two obvious corrective actions can be taken: improve the performance of the model by using more positive (genuine) and/or negative (impostor) training samples; and, in the context of fusion, which is the central topic treated in this paper, determine a subset of users who require or do not require fusion.

- *A selective fusion strategy*: Instead of using the output of all the sub-systems, we propose a fusion model that will combine only a subset of the most discriminative sub-systems, in a user-dependent manner. In this way, not all biometric sub-systems (devices) need to be operational for each transaction. The motivation of this strategy is that human can recognize people using only the most discriminatory information. This is especially true in the context of multi-modal biometrics because the participating biometric systems can operate independently of each other.

The feasibility of combining a user-ranking criterion with selective fusion strategy (hence using two of the above three mentioned strategies) has been recently supported by our work in [10]. In this study, we aim to integrate all the three mentioned strategies in a single fusion operator called the OR-switcher. As a fusion classifier, it has the following characteristics:

- user-specific, i.e., adapted to each user, and
- selective – it has a tunable criterion that chooses its constituent biometric systems to combine.

The OR-switcher is executed in three steps:

(1) Reduce the variability across users by adopting a user-specific score normalization procedure.
(2) Rank the users based on a training score dataset in such a way that the obtained ranked order will maximally correlate with the rank ordering that is obtained if it were computed on the test set.
(3) Selectively combine the sub-system outputs. If the OR-switcher determines that a single system is "good enough", then, fusion is not needed.

Due to its selective fusion capability, the OR-switcher performs fusion only when necessary, exploiting the imbalance in performance across the user models. This is made possible by a criterion to rank them.

Although selective fusion has been studied, for instance, [34], that study did not take into account the user-induced score variability.

Our contributions can be summarized as follows: (1) a thorough analysis of generalisation ability of user-specific parameters, which leads to (2) the proposal of a user ranking criterion capable of generalising to unseen data, and (3) advancement in user-specific fusion methodology via user-selective fusion. This paper goes a long way in extending the analysis of user-specific parameters as reported in [40]. First, the current paper investigates the generalisation ability of these parameters to unseen data in the context of multimodal biometric fusion. Second, we investigate in details the effect of pre-processing the biometric output in order to improve its central tendency which is central to parameterisation of any user-specific score normalisation procedure. Last but not least, a user-selective fusion strategy is proposed.

1.6. Paper organization

This paper is organized as follows: Section 2 first presents an overview of a standard multimodal benchmark database to be used. Section 3 then presents a moment-based user-specific analysis as a tool to describe Doddington's zoo/biometric menagerie. Section 4 explains the three important steps of the OR-switcher. Section 5 gives a possible implementation of the algorithm. Section 6 then presents some empirical findings carried out on the 15 XM2VTS multimodal fusion tasks. Finally, Section 7 concludes the paper. The XM2VTS score-level benchmark database will be first presented in Section 2 since it is used in Sections 3–6. A detailed description of this database can be found in [41].

2. The XM2VTS score-level benchmark database

The XM2VTS database [42] contains synchronized face video and speech data from 295 subjects, of which, 200 are classified as genuine users and 95 as impostors. Out of the 95 impostors, 70 are used in the fusion development (i.e, training) set and 25 in the fusion evaluation (test) set. The data were processed independently by multiple face and speech algorithms (sub-systems) [41] according to the Lausanne Protocols I and II resulting in the generation of match scores. The two protocols differ mainly in the way the development (training) data is partitioned to build the baseline systems. The evaluation (test) data in both protocols *remain the same*.

All speech systems are based on Gaussian mixture models (GMMs) [43] and differ only by the nature of the feature representation used: Linear Frequency Cepstral Coefficients (LFCC) [44], Phase-Auto-Correlation (PAC) [45] and Spectral Subband Centroids (SSC) [46,47]. These feature representations are selected such that they exhibit different degrees of tolerance to noise. We observe that the highly tolerant feature representation schemes perform worse in clean conditions whilst the highly accurate feature representation schemes degrade quickly under noisy conditions.

The face system considered in this study is based on the Discrete Cosine Transform (DCT) coefficients [48]. The DCT procedure operates with two image block dimensions, i.e., small (s) or big (b), and is denoted by DCTs or DCTb, respectively. Hence, the matching process is local as opposed to a holistic matching approach such as the Principal Component Analysis [49].

Table 1 presents a list of baseline experimental scores used in this study. The score data set is publicly available at "http://www.idiap.ch/~norman/fusion" and was reported in [41]. Note that each system can be characterized by a feature representation scheme and a classifier. Two types of classifiers were used, i.e., GMMs and multi-layer Perceptrons (MLPs).

The score data sets used here are very similar to [41] with the following minor changes:

- A face system based on a downsized face image concatenated with color Histogram information (FH) [50] was *not* considered here since a procedure, aimed at improving the central tendency of match scores (to be described in Section 3.3), gives $-\infty$ and ∞ values. We originally included this system in our experiments and handled these two exceptions but determined that its inclusion only complicates the analysis without giving any additional insight into the problem.

Table 1
The 13 baseline experiments taken from the XM2VTS benchmark fusion database were considered for studying the user-specific statistics as well as the proposed OR-switcher fusion operator.

Labels	Modalities	Features	Classifiers	Used in fusion
P1:1	Face	DCTs	GMM	Yes
P1:2	Face	DCTb	GMM	Yes
P1:3	Speech	LFCC	GMM	Yes
P1:4	Speech	PAC	GMM	Yes
P1:5	Speech	SSC	GMM	Yes
P1:6	Face	DCTs	MLP	No
P1:7	Face	DCTs	MLPi	Yes
P1:8	Face	DCTb	MLP	No
P1:9	Face	DCTb	MLPi	Yes
P2:1	Face	DCTb	GMM	Yes
P2:2	Speech	LFCC	GMM	Yes
P2:3	Speech	PAC	GMM	Yes
P2:4	Speech	SSC	GMM	Yes

P$m:n$ denotes the n-th system in the m-th protocol. MLPi denotes the output of MLP converted to LLR using inverse hyperbolic tangent function. P1:6 and P1:7 (resp. P1:8 and P1:9) are the *same* systems except that the scores of the latter have been transformed.

- We applied the above mentioned probabilistic to two score data sets, labeled as "P1:6" and "P1:8" in Table 1. The resulting score sets are labeled as "P1:7" and "P1:9".

While all the 13 score data sets, including the transformed ones using the above mentioned procedure, are used in the experiments reported in Section 3, the last column in Table 1 shows the actual systems used in the fusion experiments, which will be reported in Section 4. According to this column, by exhaustively pairing the available face and speech systems in both the Lausanne Protocols (noting that P1:6 and P1:8 are not used), we obtained $4\times 3+1\times 3=15$ fusion possibilities. The pairing of these 15 fusion tasks is shown in Fig. 7.

3. User-specific analysis of scores

This section will begin by introducing some notation. Then, we will present a mechanism to describe Doddington's phenomenon. The remaining sections will then analyze the zoo effect based on real data.

3.1. Notation

Let $j\in\mathcal{J}$ to be one of the identities $\{1,\dots,J\}$ and there are J users. We also assume that there is only a single model (template) associated with each user. In a usual identification setting, one compares all samples belonging to $j'\in\mathcal{J}$ against the model of j (the target user) in order to obtain a score set $\mathcal{Y}(j,j')$. When $j'=j$, the matching score set is said to be genuine, whereas when $j'\neq j$, it is said to be non-match or impostor. We further introduce two *user-specific* score sets, both dependent on the claimed identity: $\mathcal{Y}_j^{\mathtt{G}}\equiv\mathcal{Y}(j,j)$ for the genuine class, and $\mathcal{Y}_j^{\mathtt{I}}$ for the impostor class. In this study, for the impostor class, the scores are a *union* or aggregation of all other users except the target user j $\mathcal{Y}_j^{\mathtt{I}}\equiv\cup_{j'\in\mathcal{J},j'\neq j}\mathcal{Y}'(j,j')$.

3.2. Describing user-specific class-conditional score distributions using moments

Using the above notation, we shall use the score variable y to represent an element in the set $\mathcal{Y}_j^k$ for a given class $k=[\mathtt{G},\mathtt{I}\}$ (genuine or impostor) and a given claimed identity j. The unknown distribution from which $\mathcal{Y}_j^k$ was generated is denoted by $p(y|k,j)$. Thus, the unconditional distribution of y is $p(y)=\sum_{k,j}p(y|k,j)P(j|k)P(k)$ where $P(j|k)$ is the prior class-conditional probability claiming identity j and $P(k)$ is the prior class probability. Similarly, the class-conditional distribution is given by

$$p(y|k)=\sum_j p(y|k,j)P(j|k) \tag{1}$$

The user-specific class-conditional expected value of y is

$$\mu_j^k=\int_y p(y|k,j)y\,dy\equiv\mathbb{E}_y[y|k,j]$$

The global (system wide) class-conditional expected value of y is

$$\mu^k=\sum_{j=1}^{J}\left(\int_y p(y|k,j)y\,dy\right)P(j|k)$$
$$\equiv\mathbb{E}_j[\mathbb{E}_y[y|k,j]|k]=\mathbb{E}_j[\mu_j^k|k] \tag{2}$$

where $P(j|k)=P(j,k)/P(k)$ and we have used the following term

$$\mathbb{E}_j[\bullet|k]\equiv\sum_{j=1}^{J}\bullet P(j|k)$$

to denote the expectation over all users (enrollees), conditioned on k. We note here that the global mean, μ^k, is a function of user-specific mean, μ_j^k.

Let the global variance be defined as $(\sigma^k)^2\equiv\mathbb{E}_y[(y-\mu^k)^2|k]$ and $(\sigma_j^k)^2\equiv[(y-\mu_j^k)^2|k,j]$ be the user-specific variance. In the appendix, we show that these two variances are related by

$$(\sigma^k)^2=\mathbb{E}_j[(\sigma_j^k)^2|k]+\mathbb{E}_j[(\mu_j^k-\mu^k)^2|k] \tag{3}$$

We note again that the global variance is a function of user-specific variance. The dependence of the class-conditional global moments on the user-specific moments is consistent with Eq. (1) which says that the global (class-conditional) distribution is effectively a mixture of user-specific score distributions.

The moment-based analysis presented here has a very important implication on the analysis of Doddington's zoo that we will carry out on real data (to be described in Section 3.4). If one were to use only the first two orders of moment to analyze the score distributions, then these moments must be able to describe the data sufficiently well. If the distributions are skewed, the first two orders of moment will not be sufficient to describe the data. In this case, there are two solutions: use higher order moments, or pre-process the score so that only the first two orders of moment are sufficient.

In the first approach, one has to estimate the skewness of the user-specific class-conditional distributions. Due to lack of genuine score samples (only two or three in our experiments), it is often impossible to estimate variance, let alone skewness (and kurtosis). Since the first approach is not realizable, we shall use the second approach. A possible procedure to pre-process the data to exhibit high central tendency of distribution (which can adequately be described by the first two order of moments) is described in Section 3.3. In other words, we require that the *transformed* global (class-conditional) score distributions to be approximately Gaussian.

Recall that the objective of the second approach is to describe the user-specific (class-conditional) distributions using only the first two moments, and ignoring higher moments (skewness and kurtosis) that cannot be practically computed. Effectively, the unknown user-specific distribution can be approximated by a Gaussian distribution (with zero skewness and zero kurtosis w.r.t. a Gaussian). It is natural to ask why by ensuring that the global class-conditional distribution exhibit high central tendency that the user-specific distribution may be adequately be specified by the first two order of user-specific moments. An explanation of this follows from Eq. (1): the global class-conditional match scores are functionally dependent on the user-specific class-conditional match scores. However, we also know that the global class-conditional match score distribution *cannot* be Gaussian, since it is, by definition, a mixture of user-specific (class-conditional) score distributions; and the form of the latter distribution is often unknown due to the lack of samples needed to perform a standard Gaussian test. For instance, in face verification, pose and illumination can greatly affect the genuine match scores distribution. It is therefore difficult to determine the distribution in real life due to the compounded effect of these factors. This implies that seeking a perfect transformation to ensure that the system-wide scores conform to a Gaussian distribution is not the goal, but to improve its central tendency (zero skewness) is.

In this light, we took a rather practical approach in this paper, summarized in the following steps:

(1) Project the system-wide match scores to exhibit high central tendency (to be described in Section 3.3).
(2) Estimate the first two order of user-specific class-conditional moments.
(3) Rigorously test the estimates of the user-specific moments using different data sets (Refer Section 3.4).

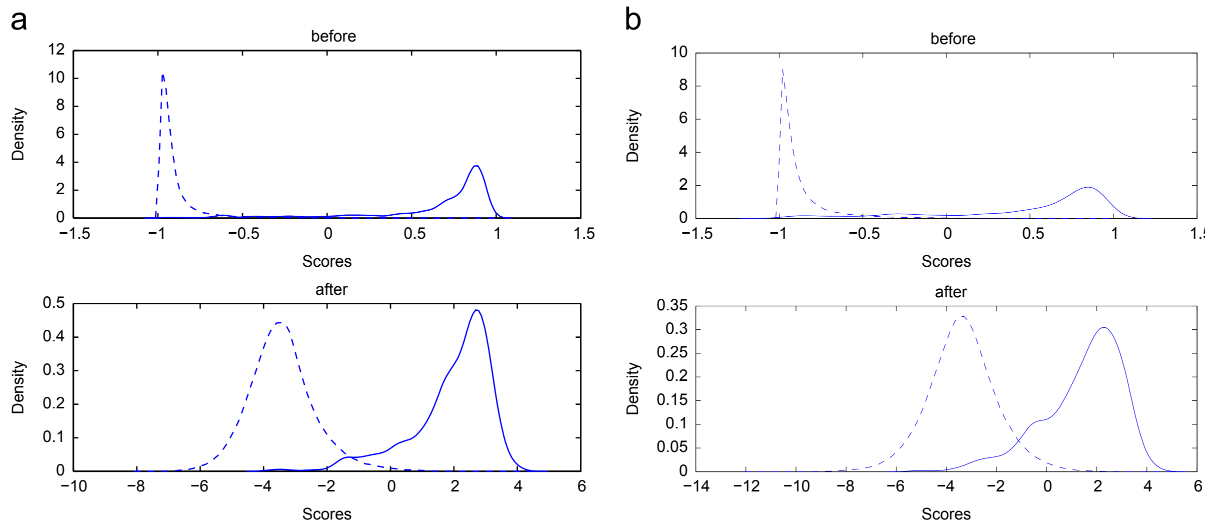

Fig. 1. Genuine (continuous lines) and impostor (dashed lines) density of match scores before (top) and after (bottom) transformation. (a) P1:6 (before) and P1:7 (after) and (b) P1:8 (before) and P1:9 (after).

To illustrate the utility of user-specific moments, we shall present some criteria that have been used successfully to measure the discriminative power of each user model using only the first two orders of user-specific moments (see Section 3.5). Finally, we also relate the user-specific moments to the Doddington's zoo phenomenon in Section 3.6 using a classifier taken from the XM2VTS score database as an example.

3.3. *Improving the central tendency of match scores*

When the output of a classifier is bounded in $[a, b]$, in [51], it was recommended that the following order-preserving transformation is used:

$$y' = \log\left(\frac{y-a}{b-y}\right) \tag{4}$$

As an example, if a classifier output corresponds to the probability of being a client given an observed biometric sample x, i.e., $y^{prob} = P(\mathbb{G}|x)$, then $a=0$ and $b=1$. The above transformation becomes

$$\begin{aligned} y^{llr} &= \log\left(\frac{y}{1-y}\right) = \log\left(\frac{P(\mathbb{G}|x)}{P(\mathbb{I}|x)}\right) \\ &= \log\left(\frac{p(x|\mathbb{G})}{p(x|\mathbb{I})}\right) + \log\left(\frac{P(\mathbb{G})}{P(\mathbb{I})}\right) \\ &= \underbrace{\log\left(\frac{p(x|\mathbb{G})}{p(x|I)}\right)} + \text{const.} \end{aligned} \tag{5}$$

We note here that the above underbraced term is another way of constructing a classifier known as the likelihood ratio test and the constant is the theoretically optimal decision threshold in this log-likelihood ratio space. There is a practical advantage of working in the log-likelihood space: its class-conditional density in the probability space is less skewed, for both the system-wide and user-specific component distributions, as opposed to the probability space.

We shall illustrate the application of Eq. (4) to two systems, i.e., P1:6 and P1:8, as already briefly mentioned in Section 2. Both these systems have output values that lie in the range of −1 and 1 due to the hyperbolic tangent function (noting that the extreme values of −1 and 1 are never observed). We therefore applied Eq. (4) to both the system outputs by setting $a=-1$ and $b=1$.[2]

[2] It can be shown that the inverse of a hyperbolic tangent function is $\text{argtanh}(y) = \frac{1}{2}\log(1+y)/(1-y)$. Hence, Eq. (4) differs from argtanh by a negligible constant factor.

The match score distributions before and after transformation are shown in Fig. 1. As can be observed, there is an obvious improvement in terms of central tendency.

In order to assess the improvement in central tendency, we propose to use the following three objective measures:

- the Komolgorov–Smirnov (KS) statistic: It is a measure of deviation from a normal distribution and is bounded in [0,1]. Smaller KS values imply better conformance to a normal distribution.
- skewness: It is a measure of the asymmetry of the data around the sample mean. If skewness is negative, the data are spread out more to the left of the mean and vice-versa for positive skewness. The skewness of the class-conditional distribution (subject to client or impostor matching) is defined by:

$$skewness = \frac{\mathbb{E}_y[y-\mu^k|k]^3}{(\sigma^k)^3}$$

  The distribution of a Gaussian has zero skewness.
- kurtosis: It is a measure of how outlier-prone a distribution is. We use the following definition of kurtosis (applied to class-conditional match scores):

$$kurtosis = \frac{\mathbb{E}_y[y-\mu^k|k]^4}{(\sigma^k)^4}$$

  As a guide, a Gaussian distribution has a kurtosis of 3.

Fig. 2 shows the three measures on the system-wide class-conditional scores whereas Fig. 3 presents similar results but conditioned on each user model and only for the impostor match scores.

The three objective measures for the genuine scores are either not computable (which is the case for the KS value) or deemed useless due to insufficient sample size, recalling that only 2 or 3 samples are available for the user-specific genuine scores. In comparison, there are $25 \times 8 = 200$ impostor samples (25 out-of-sample subjects each contributing 8 samples) for the development set and $75 \times 8 = 600$ samples for the evaluation set. In both figures, it can be observed that the KS values are reduced in *all* eight data sets (from these three dichotomies: development or evaluation, client or impostor, and, P1:6 or P1:8). The skewness and kurtosis are also reduced drastically after applying Eq. (4). The above observation is consistent with Eq. (1) showing that

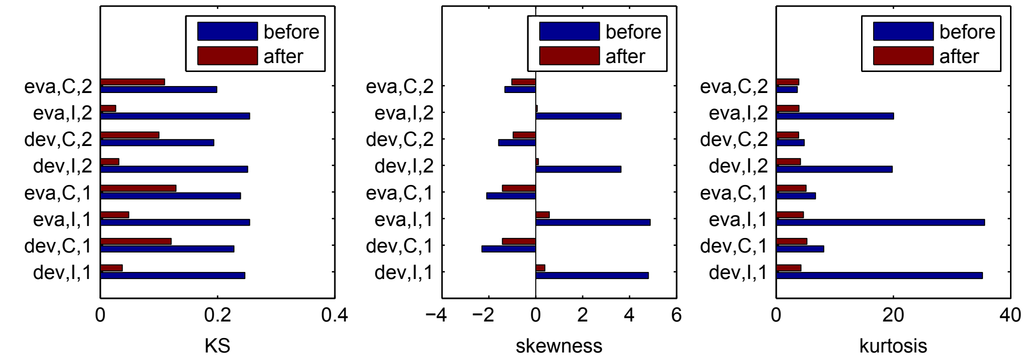

Fig. 2. Boxplots of KS value, skewness, and kurtosis of P1:6 (denoted as system 1 before transformation), P1:7 (system 1 after transformation), P1:7 (system 2 before transformation), P1:8 (system 2 after transformation) evaluated on both the development (denoted as dev) and evaluation (eva) sets of the client (denoted as C) and impostor (I) scores. A box-plot shows the 25-th and 75-th percentile of a statistic using a box, with the median value in red. The 5-th and 95-th percentile intervals are shown in dashed lines extending from the box. Finally, extreme values are plotted with "+" sign. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

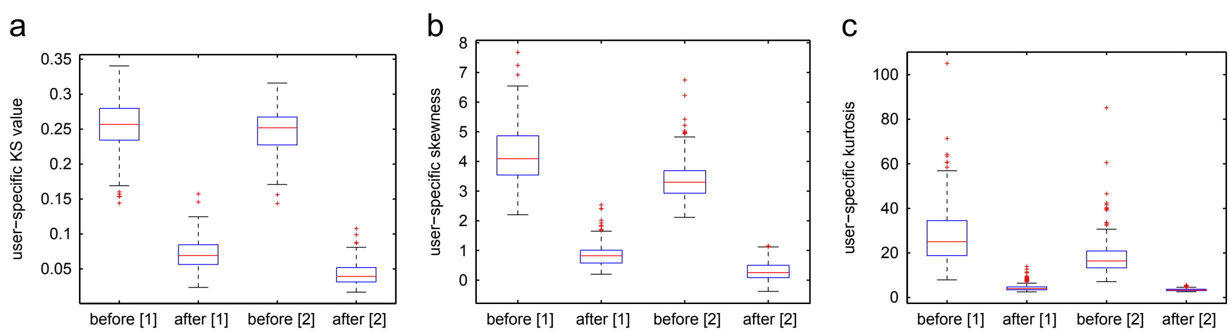

Fig. 3. Boxplots of KS value, skewness and kurtosis of P1:6 (denoted as system 1 before transformation), P1:7 (system 1 after transformation), P1:8 (system 2 before transformation), P1:9 (system 2 after transformation) evaluated on the evaluation set of the impostor scores. Similar results are obtained for the development set (not shown here). The three objective measures for the client scores are either not computable (for the KS value) or deemed useless due to insufficient sample-size (with only 2 or 3 samples). See the caption of Fig. 2 for interpreting a boxplot. (a) KS, (b) skewness and (c) kurtosis.

the system-wide score distribution is a mixture of user-specific score distributions. Because of this, the system-wide score distribution is a function of the user-specific score distribution. As a result, even if the genuine user-specific score distributions are not observable, we conjecture that, after transformation, they have reduced skewness and kurtosis, making the distribution closer to a Gaussian distribution, although not as close as their impostor counterparts. The higher normality of the user-specific impostor distributions compared to their client counterparts can be attributed to the compounded effect of impostor match scores, recalling that user-specific impostor scores is an aggregation of scores contributed by a set of out-of-sample subjects (see Section 3.1).

3.4. Predictability of user-specific statistics under session mismatch

In order to perform user-specific analysis, we have spent a great deal of effort in the preceding sections to ensure the central tendency and better conformance to normal distribution of user-specific class-conditional scores. This allows us to characterize each user-specific distribution by only the first two orders of moments. This section validates the *predictability*, or stability of these statistics in different data sets, i.e., development versus evaluation set.

Two factors can affect the predictability of user-specific statistics: cross-session matching and the computability of the estimates. The first factor is due to the fact biometric matching is an inherently stochastic process, i.e., two biometric samples can never be matched perfectly. This variability is amplified when samples are collected in different sessions (hence termed cross-session matching) in speaker verification [52]. The second factor concerns whether or not an estimate can be reliably computed due to small sample-size. For instance, the genuine scores of our data set has only 2–3 genuine sample per user model whereas larger databases such as the NIST2005 speaker evaluation database [52] have an average of 10 samples per subject. In comparison, the number of impostor scores is in the order of hundreds. As a result, we expect that the user-specific impostor statistics, μ_j^{I} and σ_j^{I}, can be estimated reliably unlike their genuine counterparts, μ_j^{G} and σ_j^{G}.

In order to quantify the predictability of user-specific statistics (for μ_j^{G} and σ_j^{G} separately), we propose to measure the correlation between the estimates of these statistics derived from the development set and the evaluation set. A perfect correlation of one indicates perfect predictability; and this is impossible to attain due to the stochastic nature of biometric matching and the acquisition process. On the other hand, a correlation of zero indicates that a given statistic cannot generalize from the development score data set to the unseen evaluation one.

This correlation-based predictability analysis was performed on the 13 XM2VTS score data sets. One of the experimental outcomes is shown in Fig. 4(a). The result of the experiments are summarized in Fig. 4(b) in boxplots. The experimental outcome shows that σ_j^{G} is not at all predictable (with correlation around 0.2) whereas μ_j^{G} is somewhat reliable (with correlation around 0.6). Furthermore, the statistics μ_j^{I} and σ_j^{I} are not sensitive to the choice of the

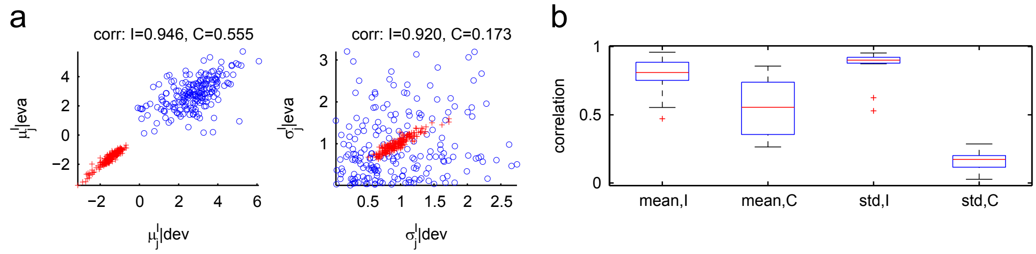

Fig. 4. (a) An example of scatter plot of $\mu_j^k|eva$ versus $\mu_j^k|dev$ and $\sigma_j^k|eva$ versus $\sigma_j^k|dev$ taken from one of the 13 score data sets. For each scatter plot, two correlation values are measured, one for user-specific impostor parameter and another for the genuine one. In (a) the left panel shows the user-specific mean parameter whereas the right one shows its corresponding standard deviation parameter. For both diagrams, '+' indicates a user-specific impostor parameter (mean or standard deviation) whereas 'o' indicates a user-specific genuine parameter. The correlation values of each scatter plot, are indicated on the top of each plot and (b) a summary of the robustness of user-specific statistics μ_j^k and σ_j^k for $k=\{\mathtt{G},\mathtt{I}\}$ across all users $j\in\mathcal{J}$ in terms of correlation of box-plots. Each bar contains 13 correlation values which correspond to the 13 score data sets. The two outliers in $\sigma_j^{\mathtt{I}}$ are due to of P1:6 (MLP,DCTs) and P1:8 (MLP,DCTb), respectively. Similarly the outlier in $\mu_j^{\mathtt{I}}$ is due to P1:6 (MLP,DCTs). These outliers show that if the scores are not Gaussian, the user-specific statistics are not useful. This experiment also confirms the effectiveness of probabilistic inversion to remove this undesired effect. (a) Scatter plots of user-specific statistics and (b) Summary using correlation.

impostor population, i.e., if two sets of casual impostors try to impersonate the same user, the user-specific impostor statistics due to the first and second impostor populations are still *strongly* correlated (about 0.8 for $\mu_j^{\mathtt{I}}$ and 0.9 for $\sigma_j^{\mathtt{I}}$). We conjecture that this observation does not necessarily apply to the case where one set of impostors is casual (zero-effort) and the other set is concerted (deliberate spoofing). This is a subject of future investigation.

3.5. Quantifying user-specific discrimination power

If $p(y|k,j)$ is normally distributed, the corresponding False Rejection Rate (FRR) and False Acceptance Rate (FAR) will each be an integral of a Gaussian function, i.e.,:

$$\mathrm{FRR}_j(\Delta)=P(y\le\Delta|\mathtt{G},j)=\frac{1}{2}+\frac{1}{2}\mathrm{erf}\left(\frac{\Delta-\mu_j^{\mathtt{G}}}{\sigma_j^{\mathtt{G}}\sqrt{2}}\right), \quad (6)$$

$$\mathrm{FAR}_j(\Delta)=1-P(y\le\Delta|\mathtt{I},j)=\frac{1}{2}-\frac{1}{2}\mathrm{erf}\left(\frac{\Delta-\mu_j^{\mathtt{I}}}{\sigma_j^{\mathtt{I}}\sqrt{2}}\right), \quad (7)$$

where

$$\mathrm{erf}(z)=\frac{2}{\sqrt{\pi}}\int_0^z \exp[-t^2]\,dt,$$

which is known as the "error function" in statistics. Similar derivations of FAR and FRR at the global level, i.e., without the user index j, can be found in [53]. By plotting $(\mathrm{FAR}_j(\Delta),\mathrm{FRR}_j(\Delta))$, for all $\Delta\in[-\infty,\infty]$, one obtains a user-specific Receiver Operating Characteristic (ROC) curve. The unique point where both FAR and FRR intersect each other, i.e., $\mathrm{FAR}_j(\Delta^*)=\mathrm{FRR}_j(\Delta^*)$, is called the Equal Error Rate (EER) and this happens at threshold

$$\Delta^*=\frac{\mu_j^I\sigma_j^G+\mu_j^G\sigma_j^I}{\sigma_j^I+\sigma_j^G} \quad (8)$$

giving:

$$\mathrm{EER}_j=\frac{1}{2}-\frac{1}{2}\mathrm{erf}\left(\frac{\mathrm{F{-}ratio}_j}{\sqrt{2}}\right), \quad (9)$$

where we introduced the user-specific F-ratio ("F" for Fisher), defined as

$$\mathrm{F{-}ratio}_j=\frac{\mu_j^G-\mu_j^I}{\sigma_j^G+\sigma_j^I}. \quad (10)$$

This term measures the class-separability of a given user. Other measures include the d-prime statistics [38] and the two-class Fisher ratio [37], defined as

$$d'_j=\frac{|\mu_j^{\mathtt{G}}-\mu_j^{\mathtt{I}}|}{\sqrt{\frac{1}{2}(\sigma_j^{\mathtt{G}})^2+\frac{1}{2}(\sigma_j^{\mathtt{I}})^2}} \quad \text{and} \quad \mathrm{Fisher{-}ratio}_j=\frac{\mu_j^{\mathtt{C}}-\mu_j^{\mathtt{I}}}{(\sigma_j^{\mathtt{G}})^2+(\sigma_j^{\mathtt{I}})^2},$$

respectively. Among these three criteria, the F-ratio has the advantage that it is related to EER in a *closed form* due to Eq. (9). In case of violation of this assumption, it can still be shown via simulation that the estimated EER can be positively or negatively biased.[3] The ranked user turns out to be robust to such bias since *all* user-specific distributions will subject to the *same distortion*, hence, will in principle, have the same bias. In [36], the Fisher ratio, F-ratio and d-prime statistics were used successfully to rank users on three biometric modalities, i.e., face, fingerprint and iris. This provides further evidence of the robustness of these statistics under deviation from the Gaussian assumption.

An impending problem in relation to the application of these criteria is that, from Section 3.4, we know that $\sigma_j^{\mathtt{G}}$ is not computable due to small sample-size. This will be treated in Section 4.2.

3.6. User-specific statistics and Doddington's menagerie

Having discussed the estimates of user-specific statistics, this section attempts to relate the Doddington's menagerie with these statistics, and their utility in reducing this phenomenon. Fig. 5 shows the match score distribution for each user model and each class (genuine or impostor) of one of the XM2VTS systems. Although there are 200 user models altogether, only 20 are shown here in order to avoid cluttering the figure. Thus, there are 20 genuine-user match score distributions and 20 impostor match score distributions. Figure (a) shows the distributions of the original scores.

Referring to Doddington's menagerie, sheep are characterized by high genuine match scores (high user-specific mean values) whereas goats are characterized by low genuine match scores (low user-specific mean). Lambs are defined as a symmetry of goats, i.e., having high impostor match scores (high user-specific impostor mean). Finally, wolves are persons who can consistently give high impostor similarity scores when matched against the user models (enrolled templates in the gallery).

Note that although we do not identify the wolves directly, as done in [2], the impostor population certainly contains some wolves. By modeling the impostor score distributions, the F-ratio

[3] See the attached supplementary materials for review.

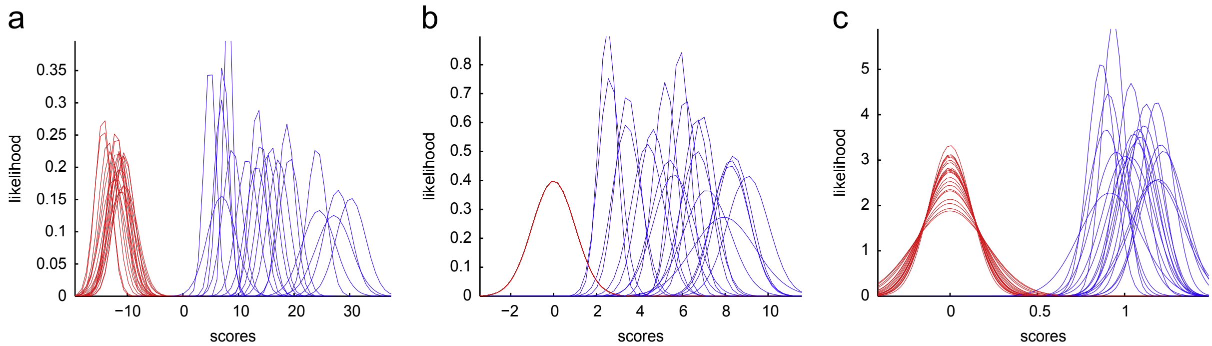

Fig. 5. Class conditional user-specific distributions of (a) a baseline system based on the P1:3 system (a GMM-based speaker verification system with LFCC features), (b) its normalized scores using Z-norm, and (c) its normalized scores using F-norm. Only the distributions of 20 users are shown here. Blue curves denote genuine score distributions; and red curves, impostor distributions. For illustration purpose, the distributions here are plotted using Gaussian and are rescaled to highlight two phenomena: the dependence of distributions on a particular user, and the effect of user-specific normalization on the resultant distributions. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

indeed takes them into consideration. What is not taken into account here is the *deliberate* or *active* impostor attack, resulting in very strong "wolves" attacking a particular enrolled subject/client. The appropriate counter measure for such attempt is to use liveness measures [54], a subject that is beyond the scope of the study of Doddington's zoo [2], which, primarily concerns separability of genuine scores from non-match impostors (in which the true identity differs from the claimed identity), i.e., in other words, quantifying the recognizability of enrolled users.

Figures (b) and (c) show the scores after applying two different user-specific score normalization procedures, called the Z-norm and the F-norm, respectively (to be explained in Section 4.1). Both normalization procedures aim to project the user-specific match scores into a common domain in such a way that a single decision threshold can be more easily found than designing one decision threshold for each user (user-specific threshold).

4. The three steps of an OR-switcher

This section describes how an OR-switcher – a user-specific and selective fusion classifier – can be built based on the three steps mentioned earlier.

4.1. Reducing user-induced variability using user-specific score normalization procedure

We examine three *primary* families of user-specific score-normalization procedures: Z-norm, F-norm and EER-norm. These three families have the following forms in the respective order:

$$y_j^Z = \frac{y - \mu_j^I}{\sigma_j^I}, \quad (11)$$

$$y_j^F = \frac{y - \mu_j^I}{\mu_j^G - \mu_j^I}, \quad (12)$$

$$y_j^{EER} = y - \Delta_j. \quad (13)$$

Eq. (11) is found in [15]; Eq. (12) in [17]; and Eq. (13) in [18] and its simplified form in [9].

The F-norm in Eq. (12) cannot be used directly because μ_j^G is stochastic in nature, as depicted in the left plot of Fig. 4(a) as well as Fig. 4(b). Two factors can be attributed to this: First, the μ_j^G obtained from an enrollment session can be different from the μ_j^G obtained from a test session. This difference is known as a cross-session variability, and can be caused by change in the environmental condition or temporary alteration of biometric traits (e.g., growth of beards, change in facial expression, etc). Second, the estimate of μ_j^G can be severely affected by the small sample-size of the user-specific genuine scores.

A practical solution to constraining the variability of μ_j^G is to compensate it with the system-wide genuine mean score, μ^G, via an adjustable parameter $\gamma \in [0, 1]$, i.e.,

$$\mu_{adj}^G = \gamma \mu_j^G + (1-\gamma)\mu^G.$$

This adaptation strategy corresponds to the maximum *a posteriori* solution where μ^G is the best guess (or the *a priori* value) of μ_j^G. By setting $\gamma = 0$, one relies only on the *a priori* value. On the other hand, by setting $\gamma = 1$ one relies entirely on the observed user-specific μ_j^G. The parameter γ can be associated with the "relevance factor" mentioned in [12]. The fundamental idea, is to further parameterize γ as a function of the number of genuine samples such that γ scales non-linearly with the sample size. However, we assume that cross validation data is not available – a realistic scenario – and so we set $\gamma = 0.5$ in all experiments. This choice has been shown to work well in our past experience [17]. The F-norm with Bayesian adaptation becomes

$$y^F = \frac{y - \mu_j^I}{\gamma \mu_j^G + (1-\gamma)\mu^G - \mu_j^I}. \quad (14)$$

In [17], the F-norm, Z-norm and EER-norm are compared. The majority of the experiments show that the F-norm is superior to the Z-norm. This can be attributed to two reasons: First, it takes the client distribution into consideration; and this information cannot be used by the Z-norm in its present form. Second, the F-norm does not rely on the second-order moment (σ_j^k for both k). As a result, it has the practical advantage of not requiring many training samples since more samples are needed as one estimates higher orders of moments. The EER-norm results in worse performance than the baseline (non-normalized system) due to the poor estimate of σ_j^G which is caused by the small sample size of user-specific genuine samples.

We shall provide an intuitive explanation why F-norm may be more advantageous than Z-norm in a situation where the user-specific genuine score distribution are *significantly different* from each other.[4] Referring back to Fig. 5, we observe that in the Z-norm domain, *all* the user-specific impostor score distributions become standard normal (with zero mean and unit variance). However, the

[4] This amount of difference can be now be quantified using a measure called the *zoo index* [55], defined as the ratio of the expected bias of mean score from a user-specific mean score (the second right hand term in Eq. (3)) and the total variance (the left hand term in Eq. (3)).

genuine-user score distributions vary significantly from one user model to another. The F-norm procedure, on the other hand, projects the match scores such that the user-specific genuine and impostor score distributions are centered on zero and one, respectively. This effect is achieved, nevertheless, with the trade-off that the user-specific impostor variances no longer become standardized as in the Z-norm domain. Empirical comparisons using various data sets [13] confirm the superiority of the F-norm over the Z-norm.

Based on above reasoning, we shall only use the F-norm as the user-specific normalization procedure of choice throughout this paper.

4.2. On user-ranking after removing user-induced variability

In the last section, we presented the F-norm as a user-specific score normalization procedure. Although such a procedure can effectively *reduce* the inter-model score variability, the mitigation is not perfect. To the authors' knowledge, there exists no procedure to remove this variability *completely*.[5] As a result, one can still compute the discrimination power of the user models in the F-norm domain. In this section, we attempt to derive a user ranking criterion *in the F-norm domain*, i.e., using scores after applying F-norm.

It is instructive to illustrate how this can be done using Fig. 5(c). In this domain, the impostor mean is always zero whereas the genuine mean is always one. Let y^F be the projected score according to Eq. (14). The first and second orders of moments of y^F are:

$$\mu_j^{F,k} = \mathrm{E}_y[y^F|k,j] \quad \text{and} \quad (\sigma_j^{F,k})^2 = \mathrm{E}_y[y^F - \mu_j^F|k,j]^2$$

where we used the super-script F to differentiate the user-specific parameters from those obtained in the original score domain (μ_j^k and σ_j^k). We note that $\mu_j^{F,\mathrm{G}} = 1$ and $\mu_j^{F,\mathcal{I}} = 0$ following observations from Fig. 5(c). This observation can also be shown mathematically:

$$\begin{aligned}
\mu_j^{F,k=\mathrm{G}} &= \mathrm{E}_y[y^F|k=\mathrm{G},j] \\
&= \mathrm{E}_y\left[\frac{y^F - \mu_j^{\mathcal{I}}}{\mu_j^{\mathrm{G}} - \mu_j^{\mathcal{I}}}\middle|k=\mathrm{G},j\right] \\
&= \frac{\mathrm{E}[y^F|k=\mathrm{G},j] - \mu_j^{\mathcal{I}}}{\mu_j^{\mathrm{G}} - \mu_j^{\mathcal{I}}} = 1
\end{aligned}$$

since $\mu_j^{\mathrm{G}} \equiv \mathrm{E}[y^F|k=\mathrm{G},j]$. The derivation for the impostor case, leading to $\mu_j^{F,\mathcal{I}} = 0$ can be done similarly.

The user-specific F-ratio in the F-norm domain becomes

$$\text{F-ratio}_j^F = \frac{\mu_j^{F,\mathrm{G}} - \mu_j^{F,\mathcal{I}}}{\sigma_j^{F,\mathrm{G}} + \sigma_j^{F,\mathcal{I}}} = \frac{1}{\sigma_j^{F,\mathrm{G}} + \sigma_j^{F,\mathcal{I}}} \tag{15}$$

However, as we have already observed in Section 3.4, the estimate of σ_j^{G} is very noisy due to small sample-size. There are two solutions to limit the variability: First, use the Bayesian adaptation strategy to compensate for the variability of $\sigma_j^{F,\mathrm{C}}$ via an adjustable parameter, γ_2,[6] i.e.,

$$(\sigma_{adj,j}^{F,\mathrm{G}})^2 = \gamma_2(\sigma_j^{F,\mathrm{G}})^2 + (1-\gamma_2)(\sigma^{F,\mathrm{G}})^2, \tag{16}$$

in order to obtain the following term:

$$\text{compensated F-ratio}_j^F = \frac{\mu_j^{F,\mathrm{G}} - \mu_j^{F,\mathcal{I}}}{\sigma_j^{F,\mathrm{G}} + \sigma_j^{F,\mathcal{I}}} = \frac{1}{\sigma_{adj,j}^{F,\mathrm{G}} + \sigma_j^{F,\mathcal{I}}} \tag{17}$$

Second, remove the term completely to obtain the following ratio, which we shall call biometric class separation ratio, or the B-ratio:

$$\text{B-ratio}_j^F = \frac{1}{\sigma_j^{F,\mathcal{I}}} \tag{18}$$

The second approach is *heuristic*, and the rationale of this is based on the observation that the estimate σ_j^{G} is not generalizable to the unseen data (see Section 3.4). Despite being a heuristic, this strategy turns out to generalise better.

Table 2 summarizes the F-ratio and its two variants applied to both the original domain and the F-norm domain.

Using the same data sets described in Section 3.4, we compared the six user-specific class-separability criteria (as listed in Table 2) for all the 13 XM2VTS score data sets derived from the development set versus its evaluation set counterpart. In this way, 13 correlation values can be measured.

Apart from using correlation, we also measured bias, which is defined as the expected value of arithmetic difference between a given criterion estimated on a development set and its counterpart estimated on an evaluation set over all users $j \in \mathcal{J}$, i.e,:

$$\text{bias} \equiv E_j[\text{F-ratio}_j|dev - \text{F-ratio}_j|eva].$$

An ideal user-specific class-separability measure should have zero bias and correlation as close to one as possible.

We summarized the correlation and bias values in Fig. 6(a) and (b) in box-plots. We also show the user-specific F-ratios of the development set versus that of the evaluation set, for all the 13 score data sets (from which a correlation value is measured), in Fig. 6(c). A similar set of figures is plotted for the B-ratio in 6(d).

As can be seen, using the original F-ratio as given, this quantity does not generalize well.

However, when the B-ratio is used, significant improvement is observed, from median correlation of 0.35–0.6. Applying B-ratio in the F-norm domain further improves the median correlation to 0.9. The high predictability of B-ratio *across different data sets* can be attributed to the effectiveness of the F-norm, coupled with the removal of the noisy term σ_j^{G}. The tight-coupling of these two techniques are not coincident. The F-norm reduces the user-specific variability by projecting all user-specific genuine means to 1 and all user-specific impostor mean to 0. The residual variability is therefore due only to the user-specific class-conditional second-order moments (variance terms) in this domain. Since the user-specific client variance cannot be estimated, describing the residual using only the user-specific impostor variance (hence B-ratio) is sufficient.

In contrast, the compensated strategy, i.e., Eq. (16). does not work as well as the B-ratio. We attempted to fine-tune γ_2 with different values and found that as this parameter is close to one, correlation improves, up to a maximum of 0.7. However, its median of bias (across the 13 XM2VTS score data sets) is still non-zero.

The experimental results clearly support that the proposed B-ratio can indeed be used to rank the user models according to their performance. This criterion can be reliably estimated from merely a *separate development* data, and at least an additional genuine sample other than the one used to build the initial template/reference model. The latter is the minimum requirement imposed by the F-norm. There are two practical implications. Firstly, one can determine the goodness of an enrollment template by using the B-ratio in conjunction with an external development database of users. When an enrollment template is judged not to be of

[5] An ideal user-specific score normalization procedure will project each user-specific class-conditional score distribution to a canonical class-conditional distribution, e.g., zero mean unit variance for the impostor distribution and unit mean unit variance for the genuine distribution.

[6] Note that this term is different from the one used in the F-norm shown in Eq. (14).

Table 2
Summary of different class-separation criteria.

Type	Original domain	F-norm domain
F-ratio	$\text{F-ratio}_j = \frac{\mu_j^C - \mu_j^I}{\sigma_j^C + \sigma_j^I}$	$\text{F-ratio}_j^F = \frac{1}{\sigma_j^{F,C} + \sigma_j^{F,I}}$
Compensated F-ratio	$\text{Compensated F-ratio}_j = \frac{\mu_j^C - \mu_j^I}{\sigma_{adj,j}^C + \sigma_j^I}$	$\text{comp. F-ratio}_j^F = \frac{1}{\sigma_{adj,j}^{F,C} + \sigma_j^{F,I}}$
B-ratio	$\text{B-ratio}_j = \frac{\mu_j^C - \mu_j^I}{\sigma_j^I}$	$\text{B-ratio}_j^F = \frac{1}{\sigma_j^{F,I}}$

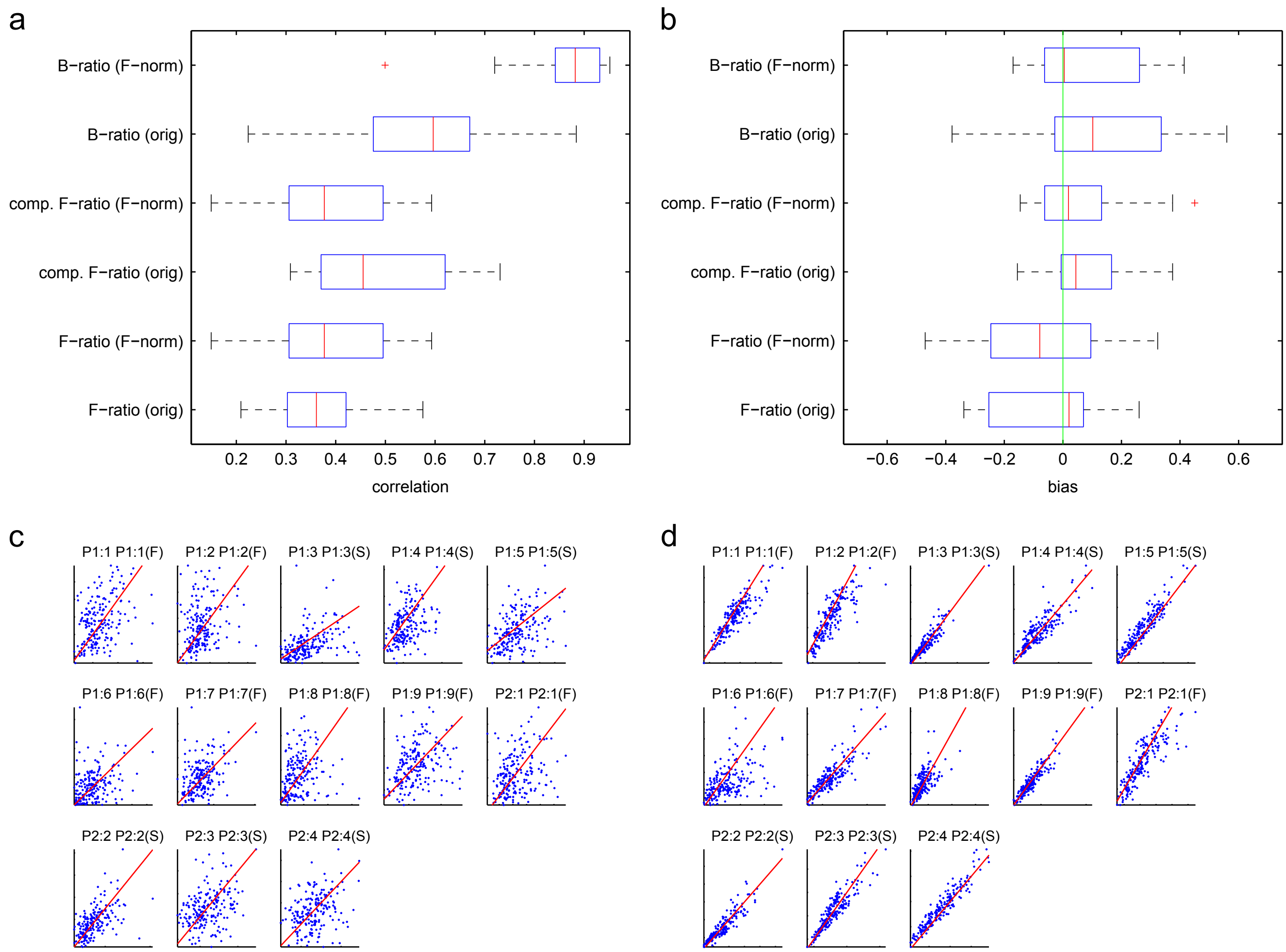

Fig. 6. Comparison of the F-ratio, compensated F-ratio (via $\gamma_2 = 0.5$ for all settings) and the B-ratio as user-specific class-separability criterion (see Table 2) across all the 13 XM2VTS score data sets. (a) Summary of the 13 experiments using correlation measures, (b) summary of the 13 experiments using bias, (c) scatter plot of the F-ratio derived from the original score domain. The *X*-axis (resp. *Y*-axis) is the F-ratio statistic calculated on the *dev* (resp. *eva*) set. Each scatter plot contains 200 points, corresponding to the 200 user models in each experiment setting and (d) Similar to (c) except that the B-ratios are used. Each diagonal (red) line in (c) and (d) represents the 45 degree line of each panel. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

sufficient quality, e.g., has a value less than a pre-determined value, corrective measures may be taken. Secondly, in the context of fusion, the B-ratio can be used to decide if fusion is indeed necessary or not. This issue is discussed in the next section.

4.3. On selective fusion

The selective fusion strategy is based on the fact that Eq. (18) can predict relative user model performance for a particular system setting. We would like to extend the user-specific B-ratio to consider the setting due to different subsets of systems p, i.e., $\text{B-ratio}_{j,p}$. For example, if there are 3 systems (hence $N=3$), p will be one of the possible power set of {1,2,3}, excluding the empty set. In our notation, we write:

$$p \in \mathcal{P}(\{1,2,3\}) - \phi \equiv \{\underbrace{\{1\},\{2\},\{3\}},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\}\}.$$

We also denote the default fusion mode that uses all the systems as $com \equiv \{1,2,3\}$. The underbraced terms are each a specific case of B-ratio_j as appeared in Eq. (18) for each of the three systems.

In order to calculate $\text{B-ratio}_{j,p}$, we first need to prepare the combined score set due to using the system subset p after applying the F-norm to each system output independently, i.e., $\{y_p^F|j\}$. Note that the F-norm has to be used because the B-ratio is only applicable to the F-normalized scores. A good candidate to use is the mean operator:

$$y_{j,p}^F = \text{mean}_{i\in p} y_{i,j}^F. \tag{19}$$

In this case, i denotes the i-th system in the set p. Since $y_{i,j}^F$ can be interpreted as an LLR, taking the sum (or mean in this case) corresponds to making the independence assumption of the system outputs $i\in p$. Using the labeled development set of scores $\{y_p^F|j,k\}$ for $k\in\{C,I\}$, we can effectively assess $\text{B-ratio}_{j,p}$ for all j and all p. Two applications are possible here. Firstly, keeping p constant and sorting $\text{B-ratio}_{j,p}$ according to j enables us to rank users for the given system subset p. Secondly, keeping j constant and sorting p for all possible combinations enables us to choose the optimal subset of sub-systems for each user. For the second application, when the system outputs $\{1,\dots,N\}$ are not correlated, e.g., in the case of multimodal fusion, the following property has to be satisfied:

$$\text{B-ratio}_{j,p2} > \text{B-ratio}_{j,p1}, \tag{20}$$

for any subset $p1, p2 \subset \{1,\dots,N\}-\varnothing$ where the number of elements in $p2$ is higher than that in $p1$, i.e., $\|p2\| > \|p1\|$. This is to say that the discriminative power of multimodal biometric increases when more biometric sub-systems are used. Unfortunately, in practice, we found that Eq. (20) is not always true because the estimated user-specific B-ratio is still unreliable enough for the second application but works fairly well for the first application. The above can be explained by the fact that $\text{B-ratio}_{j,p1}$ and $\text{B-ratio}_{j,p2}$ are not comparable when the elements in $p1$ and $p2$ are not the same. This is a commonly encountered situation in multimodal fusion where the scores produced by one classifier do not have the same statistical properties than those produced by another classifier. Furthermore, reliably estimating those parameters with a few samples would still be a difficult task, recalling that the B-ratio is estimated from samples conditioned on each user. As a result, we rank the users by $\arg\max_j$ $\text{B-ratio}_{j,p}$ where p is kept constant (instead of $\arg\max_p \text{B-ratio}_{j,p}$ by keeping j constant). Note that in this way, evaluating $\arg\max_j$ $\text{B-ratio}_{j,p}$ is possible since the system subset p is constant and hence the criterion $\text{B-ratio}_{j,p}$ (with p fixed) is comparable for different j.

5. The overall OR-switcher procedure

We have addressed the key issues of ranking users and system subsets based on the user-specific B-ratio criterion. In this Section, we propose one possible implementation of the OR-Switcher that makes use of this criterion. We will consider here the case of combining two biometric systems. The extension to N systems is straightforward. It should be noted that there are two data sets: development and evaluation sets. The development set is used to derive all the training parameters, e.g., F-norm's parameters, the user-specific and system subset dependent B-ratios, fusion classifier and the optimal decision threshold. The evaluation set is used to test the system. We utilize the following procedure:

(1) Apply the F-norm to each participating biometric system output independently. Note that the F-norm parameters must be derived from the development set.
(2) Train a Gaussian Mixture Model (GMM)-based fusion classifier of the form $y_{com}^F = \log p(\mathbf{y}^F|\mathtt{G})/p(\mathbf{y}^F|\mathtt{I})$ by estimating the class-conditional score distribution $p(\mathbf{y}^F|k)$ for each $k=\{\mathtt{G},\mathtt{I}\}$ separately using a GMM. The GMM parameters are estimated using Expectation-Maximization and its number of components are found by cross-validation [37]. In order to make the accept/reject decision, the output score $y^F com$ is compared to a threshold, which can be adjusted for different class priors.
(3) For each user $j\in\mathcal{J}$ and each possible combination subset p, assess the $\text{B-ratio}_{j,p}$ criterion given the labeled combined scores $\{y_p^F|k,j\}$ based on the development set.
(4) Sort the users in descending order based on $\text{B-ratio}_{j,com}$ (the default mode where all the systems are considered). Let r be the *pruning rate*, a parameter that is set by the system designer. This pruning rate is the proportion of users that will not require fusion. For instance, if $r=0.1$, then 10% of users will not require fusion; this means that for the remaining 90% of users fusion will always occur. In this example, the 90% of users are those whose B-ratio_{com} values are inferior, meaning that their resulting performance is likely to be inferior (hence requiring fusion). For the 10% of users, we decide the next best alternative of system subset p to use. In the case of $N=2$ systems, these alternatives are $p\in\{\{1\},\{2\}\}$. Therefore, we choose the better of the two systems, i.e.,

$$p_j^* = \arg\max_p \text{B-ratio}_{j,p}.$$

(5) During the operational phase, the combined LLR score is calculated as $y_{OR} = \log p(\mathbf{y}^F|\mathtt{G}, p_j^*)/p(\mathbf{y}^F|\mathtt{I}, p_j^*)$ where $p(\mathbf{y}^F|k, p_j^*)$ is a marginalized distribution of $p(\mathbf{y}^F|k)$ with respect to the sub-systems *not in p*. This step is performed for all users. In the case of a bimodal system, p_j^* can be $\{1\}$, $\{2\}$, or $\{1,2\}$ where 1 and 2 are the indices of the two sub-systems. Therefore, for some user, $p_j^*=\{1\}$ is considered optimal; for others, $p_j^*=\{2\}$ or $p_j^*=\{1,2\}$ (no pruning). Section 5.1 explains in detail how to marginalize a distribution estimated using a GMM.

By selective fusion, one will acquire less biometric data effectively reducing the biometric acquisition time and consequently the overall verification time. In this case, by employing the proposed selective strategy, the system can automatically utilize the most discriminatory biometric traits of an individual. The role of the B-ratio is extremely important in Step 4.

Step 4 can be omitted when $r=0$ (zero pruning rate), because in this case all the system outputs are used, and one does not need to evaluate $\text{B-ratio}_{j,p}$.

5.1. Reconciling different modes of fusion

This section describes how to compute a marginal distribution given the distribution $p(\mathbf{y}^F|k)\equiv p(\mathbf{y}^{F,k})$ estimated using GMM. Let $\mathbf{y}^{F,k} = [y_1^{F,k},\dots,y_N^{F,k}]'$ be a vector of class-conditional scores to be combined *after* applying the F-norm. Since $p(\mathbf{y}^{F,k})$ is a GMM, it can be written as

$$p(\mathbf{y}^{F,k}) = \sum_{c=1}^{N_{cmp}^k} w_c \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_c^{F,k}, \boldsymbol{\Sigma}_c^{F,k}), \tag{21}$$

where w_c is the prior of the c-th Gaussian component whose parameters are $\boldsymbol{\mu}_c^{F,k}$ and $\boldsymbol{\Sigma}_c^{F,k}$ and there are N_{comp}^k Gaussian components, for $k=\{\mathtt{G},\mathtt{I}\}$. Note that the classifier $\log p(\mathbf{y}^F|C)/p(\mathbf{y}^F|I)$ is user-independent but receives input from user-specific normalized scores obtained via the F-norm. Due to the use of the F-norm, its behavior is different for different users. In this sense, the resultant classifier is user-specific.

Given the *joint* distribution described by the mixture of Gaussian parameters $\{w_c, \boldsymbol{\mu}_c^{F,k}, \boldsymbol{\Sigma}_c^{F,k}|\forall_c\}$, our goal here is to find the marginal distribution spanned only by the subset (or subspace) $p\subseteq\{1,\dots,N\}$. One way is to marginalize the conditional joint distribution $p(\mathbf{y}^{F,k})$ with respect to the output of the systems not considered. Using the mixture of Gaussian parameters, this can be done in a rather straight-forward way. First, let us drop the super- or subscripts F, k and c from $\boldsymbol{\mu}_c^{F,k}, \boldsymbol{\Sigma}_c^{F,k}$ since the discussion that follows will always be dealing with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in the F-norm domain, applying to each k and each c Gaussian component individually. Then, the marginalized parameters due to using the subset p can be written as $\boldsymbol{\mu}_p$ and $\boldsymbol{\Sigma}_p$. The matrices before and after marginalization are related by

$$\boldsymbol{\mu} = [\boldsymbol{\mu}_p, \boldsymbol{\mu}_{\overline{p}}]'$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_p & \boldsymbol{\Sigma}_q \\ \boldsymbol{\Sigma}'_q & \boldsymbol{\Sigma}_r \end{bmatrix}$$

where $\boldsymbol{\mu}_{\overline{p}}$ is the mean vector whose elements are systems *not* in the set p; and, $\boldsymbol{\Sigma}_q$ and $\boldsymbol{\Sigma}_r$ are the rest of the sub-covariance matrices which contains the elements not in p. The above marginalization procedure for GMM can be found in [56], for instance, and is used for noisy band-limited speech recognition. Let us take an example of $N=3$ systems. Suppose the optimal subset is $p=\{1,2\}$ and the excluded system set is $\overline{p}=\{3\}$. Consequently,

$$\boldsymbol{\mu}_p = [\mu_1, \mu_2]', \quad \boldsymbol{\mu}_{\overline{p}} = [\mu_3]', \quad \boldsymbol{\Sigma}_p = \begin{bmatrix} e_{1,1} & e_{1,2} \\ e_{2,1} & e_{2,2} \end{bmatrix},$$

$$\boldsymbol{\Sigma}_q = \begin{bmatrix} e_{1,3} \\ e_{2,3} \end{bmatrix}, \quad \boldsymbol{\Sigma}_r = [e_{3,3}],$$

where $e_{m,n}$ is the m-th row and n-th column element of the covariance matrix $\boldsymbol{\Sigma}$ and $e_{m,n} = e_{n,m}$ (since the covariance matrix is symmetric). The marginalised version of GMM will be written similar to Eq 21 except that we will use $\mu^{F,k}_{p,c}$ and $\Sigma^{F,k}_{p,c}$ instead of the original parameter set.

6. Experiments

6.1. Evaluating the quality of selective fusion

Two types of evaluations are considered here, i.e., by agreement and by computational saving.

6.1.1. Evaluation by agreement

Note that p^*_j contains the subset of systems that are considered optimal, in the F-norm domain, for a user j according to the *development* set. One could evaluate the same parameter for the *evaluation* set. A useful way to evaluate if $p^*_j|dev$ is optimal or not is by comparing the same parameter derived from the evaluation set $p^*_j|eva$–which is considered the ground truth. Let $I(m,n)$ be an indicator function that outputs 1 if the sets m and n are identical and zero otherwise. The probability of choosing the "right" mode of fusion, within the population of users considered, in the OR-switcher, can be defined as

$$d = \frac{\sum_j I(p^*_j|dev, p^*_j|eva)}{J}$$

Higher d is thus clearly desired.

6.1.2. Evaluation by computational saving

One can also evaluate the computational savings due to not using some of the biometric systems. It can be quantified by

$$\text{computational saving} = 1 - \frac{\sum_{j \in \mathcal{J}} \sum_{i=1}^{N} I(\text{system}_{i,j})}{N \times J},$$

where $I(\text{system}_{i,j})$ is an indicator function that gives 1 if the i-th biometric system of user j is used and zero otherwise and there are J users. In the case of a conventional fusion classifier where all the systems are used, the computational saving is simply zero. In our case, when two systems are considered using the OR-switcher, the pruning rate r as presented in Section 5 is directly related to the computational saving in the following way:

$$\text{computational saving} = \frac{r}{N} \times 100\%.$$

For this equation, we observe that one cannot go below 50% computational saving with two systems, 33 1/3% with three systems, 25% with four systems, etc, because the maximum value r can take is 1. This means that at least one system has to be operational.

6.2. Experimental results

Using our proposed user-specific and system subset dependent B-ratio, the percentage of correctness d is measured to be 88.5% with minimum and maximum being 80% and 97.5%, respectively, across all 15 fusion tasks. This is expected since the B-ratio has the highest correlation in Fig. 6(a).

We then compared the OR-switcher with two baseline systems, as follows:

- *The conventional fusion classifier based on GMM* : In this case, the scores $\{y_i|\forall_i\}$ are used.[7]
- *The OR-switcher with various r values*: The following fraction values $r=\{0.4, 0.3, 0.2, 0.1, 0\}$ are used. When $r=0$ all the systems are used.

Fig. 7 summarizes the result of 15 fusion experiments by plotting the proposed client-specific approach as a function of the pruning rate, r. Recall that when $r=1$, no fusion is involved, e.g., only one of the two unimodal systems are selected. On the other hand, when $r=1$, both modalities are combined via the client-specific fusion. Plotted in the each sub-figure are the conventional score-level fusion system based on the GMM-Bayes classifier as well as the unimodal face and speech systems.

As can be observed, when $r=0.5$, all the client-specific systems outperform the best unimodal systems. Recall that with $r=5$, the client-specific system is composed of 50% of users utilizing unimodal systems and the remaining 50% utilizing both modalities. Furthermore, at zero pruning rate, all client-specific fusion systems outperform the baseline GMM-Bayes fusion system.

Although only EER points are presented in Fig. 7, we also assessed each of the 15 fusion tasks in DET curve.[8] An example of the second fusion task is shown in Fig. 8. In this experiment, consistent of Fig. 7(b), at $r=0$, i.e., by selectively selecting one of the two biometric modalities (hence no multimodal fusion), our user-specific strategy outperforms any single biometric system. By increasing r, the system gradually and systematically improves in performance.

6.3. Discussion

The experimental outcomes suggest that it is still possible to make decisions based on *incomplete* information. The proposed OR-switcher is really a proof of this concept. While consuming less resources (depending on the pruning rate r), the OR-switcher is at least as good as the conventional fusion classifier (in our case, a Bayesian classifier whose class-conditional densities are estimated using a mixture of Gaussian components), if not better. While it is generally true that by using a higher pruning rate the system can

[7] From our previous study [57], the GMM fusion classifier performs as well as the logistic regression and Support Vector Machines with a linear kernel. Since all these classifiers rely on the same training sets with carefully tuned hyper-parameters, their generalization performances are not expected to be *significantly* different.

[8] The DET curves for all the 15 fusion tasks can be found in "http://info.ee.surrey.ac.uk/Personal/Norman.Poh/data/expe/zoo_fusion".

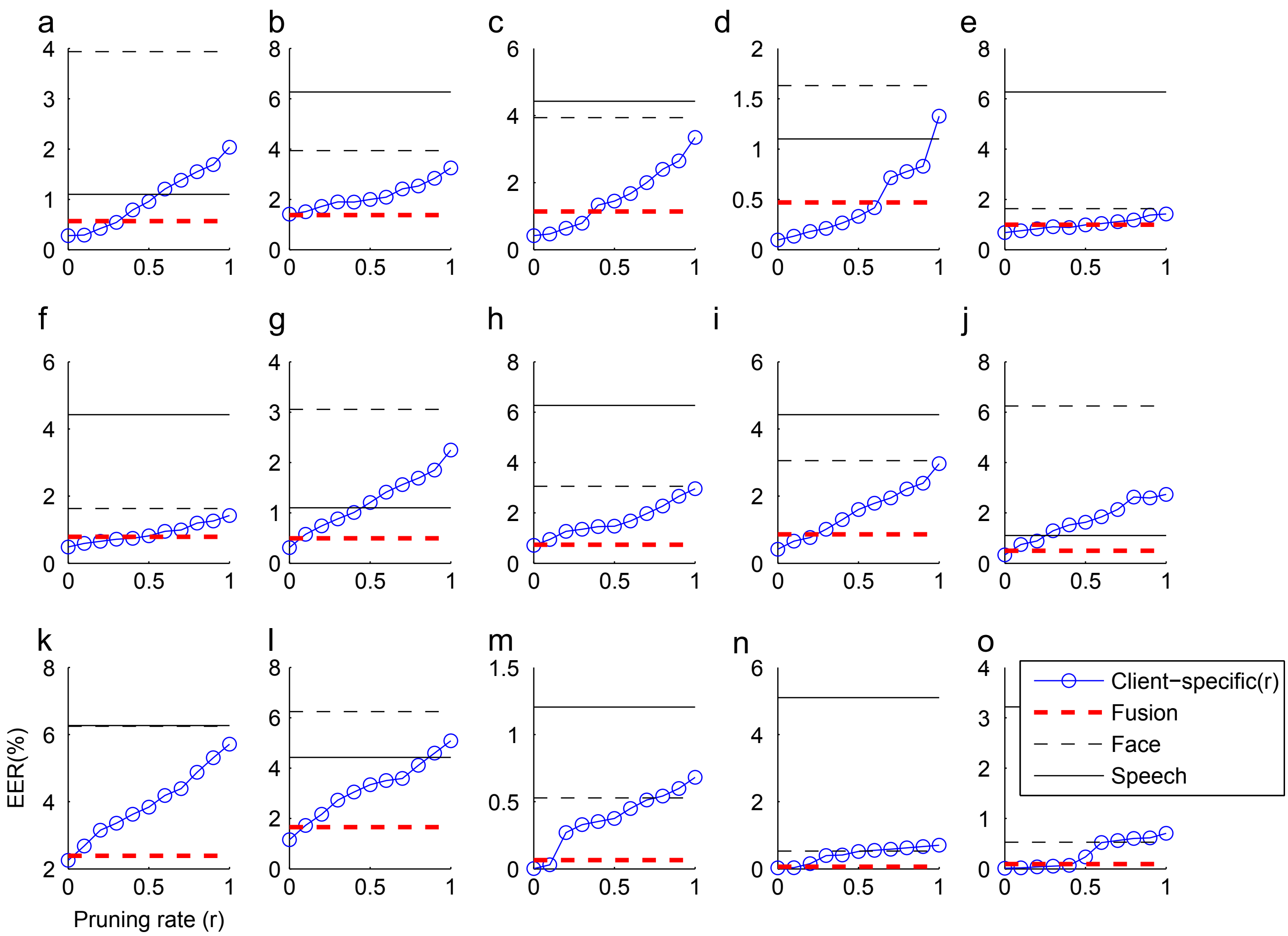

Fig. 7. Comparison of the performance of our proposed user-specific strategy as a function of the pruning rate (r) with three other systems: the baseline fusion system based on the Bayes-GMM classifier, and the unimodal face and speech systems. (a) P1:1(F)+P1:3(S), (b) P1:1(F)+P1:4(S), (c) P1:1(F)+P1:5(S), (d) P1:2(F)+P1:3(S), (e) P1:2(F)+P1:4(S), (f) P1:2(F)+P1:5(S), (g) P1:7(F)+P1:3(S), (h) P1:7(F)+P1:4(S), (i) P1:7(F)+P1:5(S), (j) P1:9(F)+P1:3(S), (k) P1:9(F)+P1:4(S), (l) P1:9(F)+P1:5(S), (m) P2:1(F)+P2:2(S), (n) P2:1(F)+P2:3(S) and (o) P2:1(F)+P2:4(S).

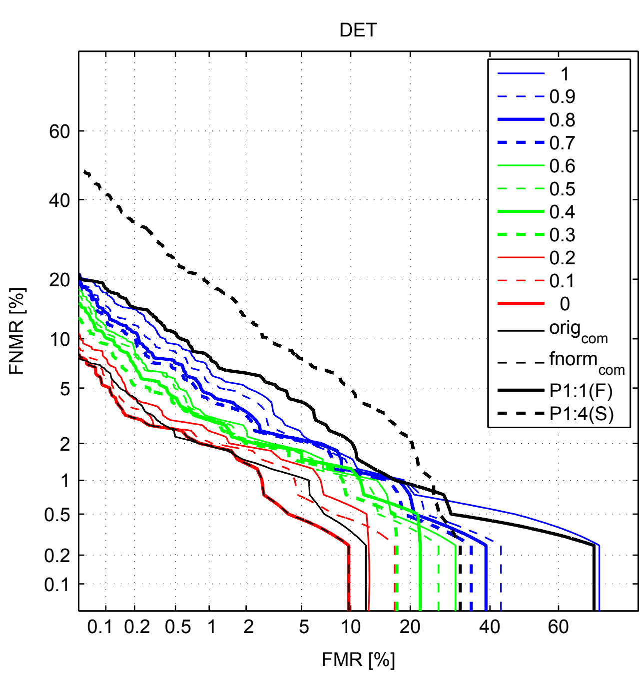

Fig. 8. An example of DET plot showing the performance of the OR-switcher with different r values, the user-independent fusion and the two face and speech unimodal systems.

degrade steadily in accuracy, the overall system also uses less resources. It is of course unfair to compare these two systems since the OR-switcher with non-zero pruning rate will always use less resources; the only fair comparison would be between two systems using the *same* amount of resources.

The added advantage of the OR-switcher is that it does not fail completely when one of its sub-systems fails to operate, as may be the case for the conventional fusion classifier. This is because the OR-switcher can capitalize on the inherent system redundancy.

In order to understand why the two-stage fusion process in the OR-switcher can perform as well as or better than the conventional (user-independent) classifier, we estimate the distribution of user-specific class-conditional scores, $p(y|k,j)$, for all k and j before normalization, as well as after normalization (using the F-norm), i.e., $p(y^F|k,j)$. For the purpose of visualization, $p(y|k,j)$ and $p(y^F|k,j)$ are each assumed to be a Gaussian with a full covariance matrix (here, we took an example of three genuine scores per user so that a full rank covariance matrix can be estimated; this is for visualization purposes only because this estimate is likely to overfit such a small amount of training data!). The Gaussian fits of model-specific class conditional scores before and after normalization are shown in Fig. 9. In (a), before the score normalization takes place, the model-specific class conditional score distributions are not well aligned. When the actual fusion performance is measured based on the pair of genuine and impostor score

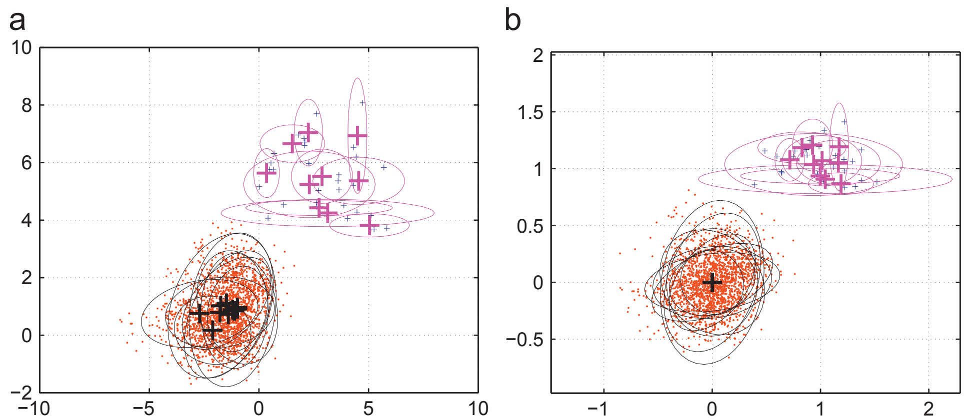

Fig. 9. An example of the effect of F-norm for both figures, the *X*- and *Y*-axes are the output score-space of a face and a speech system, respectively. The upper right clusters are client accesses whereas the lower left clusters are impostor accesses. In each figure, the scores of ten users (out of 200 available) are here; otherwise the figures would be too cluttered. This results in $10\times 2=20$ Gaussian fits, i.e., a pair of model-specific class-conditional scores for each of the ten users. (a) Before F-norm and (b) after F-norm.

distributions, it is possible to rank the users, hence, identifying goats from sheep [40]. In (b), after applying the F-norm [17], the user-specific distributions are better aligned. In this example, we show only the scores due to 20 user models selected at random. Although a perfect class separation is possible in this case, this is not so when the scores associated with all 200 user models are considered. A special property of the F-norm is that the expected model-specific genuine and impostor means *after* score normalization are 1 and 0, respectively.

7. Summary and conclusions

Following the observations in [15,2] that each user exhibits different performance in biometric authentication, we showed that it is possible to derive a criterion to rank users according to their performance strength. Such a criterion has to (a) be derived from extremely few user-specific genuine samples; (b) be able to generalize to previously unseen data; and (c) be unbiased. Guided by some preliminary experiments, we found that such a criterion is best imposed using the user-specific F-ratio after applying the F-norm. The resultant criterion is known as the B-ratio. We demonstrated the usefulness of this criterion in the context of multimodal fusion as a “switch”, and the resultant fusion operator is hence termed the “OR-swticher”. This fusion operator is different from the state-of-the-art approach because it is *user-specific*, i.e., it varies across users, and *selective*, i.e., it actively chooses only the most discriminative subsystem outputs (depending on a pre-defined pruning rate). We show that the OR-switcher outperforms some conventional (user-independent) fusion classifiers at zero pruning rate. Although increasing the pruning rate will degrade the system performance, the proposed OR-switcher is still operational and, in some cases, can perform as good as the state-of-the-art fusion classifier. Experimental results confirm our hypothesis that the OR-switcher is an effective method to handle user-induced variability in the context of multimodal biometrics.

There remains the following important concerns which need to be further investigated/validated:

- The impact of extrinsic factors, e.g., change in acquisition conditions and cross-session variability, on Doddington's menagerie and the B-ratio.
- The impact of casual and concerted impostors on Doddington's menagerie.
- The application of the B-ratio as a measure of goodness of a reference model during enrollment.
- The use of non-parametric user-specific discriminatory criterion. All the criteria studied here require the explicit use of moments and implicit Gaussian assumption on the score distributions. A possible extension to this study is to use non-parametric discriminatory criteria.

Conflict of interest statement

None declared.

Acknowledgments

This work was supported partially by the advanced researcher fellowship PA0022 121477 of the Swiss National Science Foundation and by the EU-funded Mobio project (www.mobioproject.org) grant IST-214324 and the BEAT project (no. 284989).

Appendix A

This section shows how class-conditional global (user-independent) second-order moments of match scores are *functionally* related to user-specific ones. To do so, we first calculate the class-conditional global score variance, for a given claimed identity $j=j_*$, as follows:

$$\begin{aligned}\mathbb{E}_y[(y-\mu^k)^2|k,j_*] &= \mathbb{E}_y[(y-\mu^k_{j_*}+\mu^k_{j_*}-\mu^k)^2|k,j_*]\\ &= \mathbb{E}_y[(y-\mu^k_{j_*})^2|k,j_*]\\ &= \mathbb{E}_y[(\mu^k_{j_*}-\mu^k)^2|k,j_*]\\ &\quad \underbrace{+2(\mu^k_{j_*}-\mu^k)\mathbb{E}_y[(y-\mu^k_{j_*})|k,j_*]}\end{aligned}$$

where we introduced the term $\mu^k_{j_*}$. Under the expectation $\mathbb{E}_y[\bullet|k,j_*]$, the second term is invariant, and the third (underbraced) term vanishes. As a result, we can rewrite the above equation as

$$\begin{aligned}\mathbb{E}_y[(y-\mu^k)^2|k,j_*] &= \underbrace{\mathbb{E}_y[(y-\mu^k_j)^2|k,j_*]}\\ &= (\mu^k_j-\mu^k)^2 \qquad (22)\end{aligned}$$

We recognize that the above underbraced term is the second-order moment of user-specific scores, i.e.,

$$(\sigma^k_{j_*})^2=\mathbb{E}_y[(y-\mu^k_j)^2|k,j_*] \qquad (23)$$

The global class-conditional score variance, independent of any user j, can be calculated by taking the expectation of $\mathbb{E}_j[\bullet|k]$ on both sides of Eq. (22). This is obvious by doing so on the left hand side of Eq. (22):

$$\mathbb{E}_j[\mathbb{E}_y[(y-\mu^k)^2|k,j]|k] = \mathbb{E}_y[(y-\mu^k)^2|k]$$

which is the global variance $(\sigma^k)^2$. The result of taking $\mathbb{E}_j[\bullet|k]$ on both sides of Eq. (22) can thus be written as

$$(\sigma^k)^2 = \mathbb{E}_j[(\sigma_j^k)^2|k] + \mathbb{E}_j[(\mu_j^k-\mu^k)^2|k] \tag{24}$$

References

[1] A. Ross, K. Nandakumar, A. Jain, Handbook of Multibiometrics, Springer Verlag, 2006.

[2] G. Doddington, W. Liggett, A. Martin, M. Przybocki, D. Reynolds, Sheep, goats, lambs and wolves: a statistical analysis of speaker performance in the NIST 1998 Speaker Recognition Evaluation, in: International Conference on Spoken Language Processing (ICSLP), Sydney, 1998.

[3] N. Yager, T. Dunstone, Worms, chameleons, phantoms and doves: new additions to the biometric menagerie, in: 2007 IEEE Workshop on Automatic Identification Advanced Technologies, June 2007, pp. 1–6.

[4] A. Jain, A. Ross, Learning user-specific parameters in multibiometric system, in: Proceedings of the International Conference of Image Processing (ICIP 2002), New York, 2002, pp. 57–70.

[5] R. Snelick, U. Uludag, A. Mink, M. Indovina, A. Jain, Large scale evaluation of multimodal biometric authentication using state-of-the-art systems, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (3) (2005) 450–455.

[6] J. Fierrez-Aguilar, D. Garcia-Romero, J. Ortega-Garcia, J. Gonzalez-Rodriguez, Exploiting general knowledge in user-dependent fusion strategies for multimodal biometric verification, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 5, Montreal, 2004, pp. 617–620.

[7] J. Fierrez-Aguilar, D. Garcia-Romero, J. Ortega-Garcia, J. Gonzalez-Rodriguez, Bayesian adaptation for user-dependent multimodal biometric authentication, Pattern Recognition 38 (2005) 1317–1319.

[8] A. Kumar, D. Zhang, Integrating palmprint with face for user authentication, in: Workshop on Multimodal User Authentication (MMUA 2003), Santa Barbara, 2003, pp. 107–112.

[9] K.-A. Toh, X. Jiang, W.-Y. Yau, Exploiting global and local decision for multimodal biometrics verification, IEEE Transactions on Signal Processing 52 (October (10)) (2004) 3059–3072.

[10] A. Ross, A. Rattani, M. Tistarelli, Exploiting the doddington zoo effect in biometric fusion, in: Proceedings of the 3rd IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), Washington DC, 2009.

[11] V.N. Vapnik, Statistical Learning Theory, Springer, 1998.

[12] D.A. Reynolds, T. Quatieri, R. Dunn, Speaker verification using adapted Gaussian mixture models, Digital Signal Processing 10 (1–3) (2000) 19–41.

[13] N. Poh, J. Kittler, Incorporating variation of model-specific score distribution in speaker verification systems, IEEE Transactions on Audio, Speech and Language Processing 16 (3) (2008) 594–606.

[14] A. Jain, K. Nandakumar, A. Ross, Score normalisation in multimodal biometric systems, Pattern Recognition 38 (12) (2005) 2270–2285.

[15] S. Furui, Cepstral analysis for automatic speaker verification, IEEE Transactions on Acoustic, Speech and Audio Processing/IEEE Transactions on Signal Processing 29 (2) (1981) 254–272.

[16] R. Auckenthaler, M. Carey, H. Lloyd-Thomas, Score normalization for text-independent speaker verification systems, Digital Signal Processing (DSP) Journal 10 (2000) 42–54.

[17] N. Poh, S. Bengio, F-ratio client-dependent normalisation on biometric authentication tasks, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Philadelphia, 2005, pp. 721–724.

[18] J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, Target dependent score normalisation techniques and their application to signature verification, in: Lecture Notes in Computer Science, vol. 3072, International Conference on Biometric Authentication (ICBA), Hong Kong, 2004, pp. 498–504.

[19] N. Poh, J. Kittler, On the use of log-likelihood ratio based model-specific score normalisation in biometric authentication, in: Lecture Notes in Computer Sciene, vol. 4542, IEEE/IAPR Proceedings of the International Conference on Biometrics (ICB'07), Seoul, 2007, pp. 614–624.

[20] A. Kumar, D. Zhang, Improving biometric authentication performance from the user quality, IEEE Transactions on Instrumentation and Measurement 59 (March (3)) (2010) 730–735.

[21] N. Poh, A. Rattani, M. Tistarelli, J. Kittler, Group-specific score normalization for biometric systems, in: IEEE Computer Society Workshop on Biometrics (CVPR), 2010, pp. 38–45.

[22] N. Poh, M. Tistarelli, Customizing biometric authentication systems via discriminative score calibration, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[23] A. Merati, N. Poh, J. Kittler, User-specific cohort selection and score normalization for biometric systems, IEEE Transactions on Information Forensics and Security 7 (August (4)) (2012) 1270–1277.

[24] N. Poh, J. Kittler, A unified framework for biometric expert fusion incorporating quality measures, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (January (1)) (2012) 3–18.

[25] N. Poh, A. Merati, J. Kittler, Heterogeneous information fusion: a novel fusion paradigm for biometric systems, in: International. Joint Conference on Biometrics (IJCB), 2011, pp. 1–8.

[26] Y. Fu, Z. Ma, M. Qi, J. Li, X. Li, Y. Lu, A novel user-specific face and palmprint feature level fusion, in: Second International Symposium on Intelligent Information Technology Application, 2008. IITA '08, vol. 3, IEEE, December 2008, pp. 296–300. [Online]. Available: ⟨http://dx.doi.org/10.1109/IITA.2008.469⟩.

[27] A. Wald, Sequential tests of statistical hypotheses, Annal of Mathematical Statistics 16 (1945).

[28] S. Madhvanath, V. Govindaraju, Serial classifier combination for handwritten word recognition, in: Proceedings of the Third International Conference on Document Analysis and Recognition, 1995, vol. 2, August 1995, pp. 911–914.

[29] E. Alpaydin, C. Kaynak, Cascading classifiers, Kybernetika 16 (1998) 369–374.

[30] P. Pudil, J. Novovicova, S. Blaha, J. Kittler, Multistage pattern recognition with reject option, in: 11th IAPR International Conference on Pattern Recognition, vol. II, Conference B: Pattern Recognition Methodology and Systems, Proceedings, August–3 September 1992, pp. 92–95.

[31] N. Poh, J. Kittler, On using error bounds to optimize cost-sensitive multimodal biometric authentication, in: Proceedings of the 19th International Conference on Pattern Recognition (ICPR), 2008, pp. 1–4.

[32] L.P. Cordella, P. Foggia, C. Sansone, F. Tortorella, M. Vento, A cascaded multiple expert system for verification, in: Multiple Classifier Systems, Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2000, vol. 1857, pp. 330–339. [Online]. Available: http://dx.doi.org/10.1007/3-540-45014-9_32.

[33] G. Marcialis, F. Roli, Serial fusion of fingerprint and face matchers, in: M. Haindl, J. Kittler, F. Roli (Eds.), Multiple Classifier Systems, Lecture Notes in Computer Science, vol. 4472, Springer, Berlin, Heidelberg, 2007, pp. 151–160. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-72523-7_16.

[34] L. Allano, B. Dorizzi, S. Garcia-Salicetti, Tuning cost and performance in multibiometric systems: a novel and consistent view of fusion strategies based on the Sequential Probability Ratio Test (SPRT), Pattern Recognition Letters 31 (July (9)) (2010) 884–890. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2010.01.028.

[35] K. Takahashi, M. Mimura, Y. Isobe, Y. Seto, A secure and user-friendly multimodal biometric system, 2004. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=843447.

[36] N. Poh, J. Kittler, A methodology for separating sheep from goats for controlled enrollment and multimodal fusion, in: Proceedings of the 6th Biometrics Symposium, Tampa, 2008, pp. 17–22.

[37] C. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1999.

[38] J. Daugman, Biometric Decision Landscapes, Technical Report TR482, University of Cambridge Computer Laboratory, 2000.

[39] A. Hicklin, B. Ulery, The myth of goats: how many people have fingerprints that are hard to match?, Technical Report NISTIR 7271, National Institute of Standards and Technology, 2005.

[40] N. Poh, S. Bengio, A. Ross, Revisiting doddington's zoo: a systematic method to assess user-dependent variabilities, in: Workshop on Multimodal User Authentication (MMUA 2006), Toulouse, 2006.

[41] N. Poh, S. Bengio, Database, protocol and tools for evaluating score-level fusion algorithms in biometric authentication, Pattern Recognition 39 (February (2)) (2005) 223–233.

[42] J. Matas, M. Hamouz, K. Jonsson, J. Kittler, Y. Li, C. Kotropoulos, A. Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, J. Begun, N. Capdevielle, W. Gerstner, S. Ben-Yacoub, Y. Abdeljaoued, E. Mayoraz, Comparison of face verification results on the XM2VTS database, in: Proceedings of the 15th International Conference Pattern Recognition, vol. 4, Barcelona, 2000, pp. 858–863.

[43] D.A. Reynolds, Automatic speaker recognition using Gaussian mixture speaker models, The Lincoln Laboratory Journal 8 (2) (1995) 173–192.

[44] L. Rabiner, B.-H. Juang, Fundamentals of Speech Recognition, Oxford University Press, 1993.

[45] S. Ikbal, H. Misra, H. Bourlard, Phase Auto-Correlation (PAC) derived robust speech features, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03), Hong Kong, 2003, pp. 133–136.

[46] K.K. Paliwal, Spectral subband centroids features for speech recognition, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 2, Seattle, 1998, pp. 617–620.

[47] N. Poh, C. Sanderson, S. Bengio, An investigation of spectral subband centroids for speaker authentication, in: Lecture Notes in Computer Sciene, vol. 3072, International Conference on Biometric Authentication (ICBA), Hong Kong, 2004, pp. 631–639.

[48] C. Sanderson, K. Paliwal, Fast features for face authentication under illumination direction changes, Pattern Recognition Letters 24 (14) (2003) 2409–2419.

[49] M. Turk, A. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience 3 (1) (1991) 71–86.

[50] S. Marcel, S. Bengio, Improving face verification using skin color information, in: Proceedings of the 16th International Conference on Pattern Recognition, Quebec, 2002.

[51] S.C. Dass, Y. Zhu, A.K. Jain, Validating a biometric authentication system: sample size requirements, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (12) (2006) 1319–1902.

[52] A. Martin, M. Przybocki, J.P. Campbell, The NIST Speaker Recognition Evaluation Program, Springer, 2005. (Chapter 8).
[53] N. Poh, S. Bengio, Why do multi-stream, multi-band and multi-modal approaches work on biometric user authentication tasks?, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, 2004, vol. V, pp. 893–896.
[54] M. Drahansky, D. Lodrova, Liveness detection for biometric systems based on papillary lines, International Journal of Security and its Applications 2 (4) (2008) 29–38.
[55] N. Poh, J. Kittler, A biometric menagerie index for characterising template/model-specific variation, in: Proceedings of the 3rd International Conference on Biometrics, Sardinia, 2009, pp. 816–827.
[56] A.C. Morris, M.P. Cooke, P.D. Green, Some solutions to the missing features problem in data classification with application to noise robust automatic speech recognition, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Seattle, 1998, pp. 737–740.
[57] N. Poh, S. Bengio, A study of the effects of score normalisation prior to fusion in biometric authentication tasks, IDIAP, IDIAP Research Report 69, 2004.

Norman Poh joined the Department of Computing as a Lecturer in August 2012. He received the Ph.D. degree in computer science in 2006 from the Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland. Prior to the current appointment, he was a Research Fellow with the Centre for Vision, Speech, and Signal Processing (CVSSP) and a research assistant at IDIAP research institute. His research objective is to advance pattern recognition techniques with applications to biometrics and healthcare informatics. In these two areas, he has published more than 70 publications, which also include five award-winning papers (AVBPA'05, ICB'09, HSI 2010, ICPR 2010 and Pattern Recognition Journal 2006). Other of his significant achievements include two personal research grants from the Swiss National Science Foundation (Young Prospective Researcher Fellowship and Advanced Researcher Fellowships), and Researcher of the Year 2011 Award, University of Surrey. He is a member of IEEE and IAPR, an IEEE Certified Biometrics Professional and trainer, and a member of the Education Committee of the IEEE Biometric Council.

Arun Ross received the B.E. (Hons.) degree in computer science from BITS, Pilani (India) in 1996, and the M.S. and Ph.D. degrees in computer science and engineering from Michigan State University in 1999 and 2003, respectively. Between 1996 and 1997, he was with Tata Elxsi (India) Ltd., Bangalore. He also spent three summers (2000–2002) at Siemens Corporate Research, Inc., Princeton working on fingerprint recognition algorithms. He is currently an Associate Professor in the Lane Department of Computer Science and Electrical Engineering at West Virginia University. His research interests include pattern recognition, classifier fusion, machine learning, computer vision and biometrics. He is the co-author of “Handbook of Multibiometrics” and co-editor of “Handbook of Biometrics”. He is an Associate Editor of the IEEE Transactions on Image Processing and the IEEE Transactions on Information Forensics and Security. Arun is a recipient of NSF's CAREER Award and was designated a Kavli Frontier Fellow by the National Academy of Sciences in 2006.

Weifeng Li received the M.E. and Ph.D. degrees in Information Electronics at Nagoya University, Japan in 2003 and 2006, respectively. He joined the Idiap Research Institute, Switzerland in 2006, and in 2008 he moved to Swiss Federal Institute of Technology, Lausanne (EPFL), Switzerland as a research scientist. Since 2010 he has been an associate professor in the Department of Electronic Engineering/Graduate School at Shenzhen, Tsinghua University, China. His research interests span the areas of audio and visual signal processing, Biometrics, Human-Computer Interactions (HCI), and machine learning techniques. He is a member of the IEEE and IEICE.

Josef Kittler received the B.A., Ph.D., and D.Sc. degrees from the University of Cambridge in 1971, 1974, and 1991, respectively. He heads the Centre for Vision, Speech and Signal Processing at the Faculty of Engineering and Physical Sciences, University of Surrey. He teaches and conducts research in the subject area of Machine Intelligence, with a focus on Biometrics, Video and Image Database retrieval, Automatic Inspection, Medical Data Analysis, and Cognitive Vision. He published a Prentice-Hall textbook on Pattern Recognition: A Statistical Approach and several edited volumes, as well as more than 600 scientific papers, including in excess of 170 journal papers. He serves on the Editorial Board of several scientific journals in Pattern Recognition and Computer Vision. He became Series Editor of Springer Lecture Notes on Computer Science in 2004. He served as President of the International Association for Pattern Recognition 1994–1996. Prof. Kittler was elected Fellow of the Royal Academy of Engineering in 2000. In 2006, he was awarded the KS Fu Prize from the International Association in 2006, for outstanding contributions to pattern recognition. He received the Honorary Doctorate from the Czech Technical University in Prague in 2007 and the IET Faraday Medal in 2008.