www.ietdl.org

Published in IET Biometrics
Received on 6th September 2013
Revised on 6th December 2013
Accepted on 16th January 2014
doi: 10.1049/iet-bmt.2013.0066

ISSN 2047-4938

Score calibration in face recognition

Miranti Indar Mandasari[1], *Manuel Günther*[2], *Roy Wallace*[2,3], *Rahim Saeidi*[1,4], *Sébastien Marcel*[2], *David A. van Leeuwen*[1]

[1]*Centre for Language and Speech Technology, Radboud University Nijmegen, Nijmegen, The Netherlands*
[2]*Biometrics Group, Idiap Research Institute, Martigny, Switzerland*
[3]*Zap Technology, Brisbane, Australia*
[4]*Speech and Image Processing Unit, School of Computing, University of Eastern Finland, Joensuu, Finland*
E-mail: miranti.indar.mandasari@gmail.com

Abstract: An evaluation of the verification and calibration performance of a face recognition system based on inter-session variability modelling is presented. As an extension to calibration through linear transformation of scores, categorical calibration is introduced as a way to include additional information about images for calibration. The cost of likelihood ratio, which is a well-known measure in the speaker recognition field, is used as a calibration performance metric. The results obtained from the challenging mobile biometrics and surveillance camera face databases indicate that linearly calibrated face recognition scores are less misleading in their likelihood ratio interpretation than uncalibrated scores. In addition, the categorical calibration experiments show that calibration can be used not only to improve the likelihood ratio interpretation of scores, but also to improve the verification performance of a face recognition system.

1 Introduction

Face is one of the common biometric modalities that is used by humans to perform person recognition [1]. Owing to the advancements in audio–visual recording equipment, in recent years cameras are used regularly in our everyday life. Taking photos or videos of people have become popular as camera technology for mobile devices (e.g. smart phones and tablets) rapidly improved. In the security sector, surveillance cameras are often used to monitor public places such as train stations, airports, shopping malls and hospitals. The availability of digital images from these cameras has stimulated the development of technologies to process them. One of these technologies is automatic face recognition, that is, a technology to recognise a person's identity from his or her facial image [2].

Automatic face recognition in biometrics has applications that can be divided into three main groups: commercial, governmental and forensic applications [3]. An example of commercial face recognition is the user authentication process that is performed by mobile devices and personal computers. In governmental applications, automatic face recognition systems may be used in biometric passport verification or border control activities. For both commercial and government-related applications, the subjects usually cooperate with the system. In forensic applications, digital image evidence can be recovered from surveillance operations that often involve closed circuit television (CCTV) cameras. In contrast to commercial applications, subjects in forensic face recognition generally do not cooperate with the system while such evidence is captured. Rather, they are either unaware of the system or are deliberately uncooperative, for example by hiding or disguising themselves with hats, sunglasses or masks.

Sometimes, crime scenes are watched by eyewitnesses, who may later be called upon to identify suspects. One problem of eyewitnesses is that their memory can be influenced by misleading information presented after the crime occurred [4, 5]. In cognitive psychology, this effect is called the misinformation effect paradigm [6]. Therefore, eyewitness testimonies should not be taken as the only source of information to decide whether or not the suspect is the perpetrator.

When a crime scene is monitored by a CCTV camera, the captured images are commonly compared to facial images from potential suspects of the crime by forensic experts. On one hand, humans tend to perform better than an automatic-based system when recognising familiar faces [7, 8], but on the other hand, it has been shown that automatic face recognition systems surpass human performance when comparing unfamiliar faces in difficult illumination conditions [9]. Hence, automatic systems for forensic face recognition should be used to assist forensic experts.

Several challenges emerge when images captured from mobile devices or CCTV cameras are used for face recognition. The issues that influence recognition performance include low resolution in the captured images, the pose of the subject, partial occlusions of the subject's face and variable illumination [10]. To address these issues, various techniques have been developed, including image preprocessing to reduce illumination effects [11], feature normalisation [12, 13] and inter-session variability (ISV) modelling [14]. Score normalisation techniques, such as zero and test score normalisation (ZT-norm), have also been shown to improve verification performance [15].

Generally, automatic face recognition systems compute a similarity score between a given probe sample and a model from a known identity. In authentication or verification applications of automatic face recognition, this score is compared to a threshold to classify the trial as either a client or an impostor. In forensic applications, interpreting the score is more complicated because legal decisions cannot be made directly by the automatic face comparison system but rather should be made by a judge or jury in court, after integrating information including several pieces of evidence. If the outcome of the face comparison should be presented in court, a favourable way to express it is in the form of a likelihood ratio (LR), that is, a relative likelihood of the following two competing hypotheses [16]: (a) the probe image (e.g. from CCTV) came from the suspect (prosecution hypothesis H_{P}) or (b) it originated from someone else (defense hypothesis H_{D}). It is reported that uncalibrated LRs can be misleading in their interpretation for forensics application [17, 18]. The approach that can be taken to tackle this issue is calibration [17, 19], a process to transform raw scores computed by automatic face recognition systems into calibrated LR scores.

In the field of speaker recognition, calibration is used in the speaker recognition evaluation (SRE) that is regularly held by the American National Institute for Standards and Technology (NIST) to verify advances of the technology for speaker detection systems and measuring its performance [20]. In other forensic biometric fields such as fingerprint, earmarks and signature recognition, calibration is used to transform raw scores from biometric systems to LRs [21–23]. To our knowledge, there is only limited literature available that discusses calibration for scores produced by automatic face recognition systems [21, 24].

In the previous works on face recognition, we proposed a session variability reduction method through ISV modelling [14], and a score normalisation technique via ZT-norm implementation [15] to the face recognition system. These works only focus on improving the system verification performance. Unlike the previous works, in this study, we also focus on the calibration performance and introducing calibration techniques for face recognition systems. Experiments are carried out using a face recognition system based on ISV modelling, with and without ZT-norm, and on two challenging facial image databases: mobile biometrics (MOBIO) and surveillance camera face (SCface). We evaluate both the verification and calibration performances, before and after the linear calibration is applied to the scores. We then introduce categorical calibration as a way to utilise additional information about facial images for calibration. With categorical calibration, we show that not only calibration, but also the verification performance can be improved. In the discussion, we examine the effects of calibration on score distributions produced by the face recognition system.

One important aspect of the research in this paper is that we provide the source code for all experiments, evaluations, tables and plots that are shown in Section 7. All experiments solely rely on open source software and are, therewith, entirely reproducible.

The remainder of this paper is structured as follows: the face recognition system is explained in more detail in Section 2, followed by introduction of LR calibration in Section 3 and metrics used to evaluate the system performance in Section 4. In Section 5, we present databases and evaluation protocols. The experimental setup is detailed in Section 6. Finally, the results of all experiments are discussed in Section 7, and Section 8 concludes the paper.

2 Face recognition

Automatic face recognition is the task of recognising people from their facial images. There are several challenges that influence automatic face recognition systems, like facial expressions, different illumination conditions, partial occlusions of the face, non-frontal pose and low image resolution.

Before the person shown in an image can be identified, the face has to be detected. Since we want to investigate face recognition, rather than face detection, we use the hand-labelled eye positions that are provided with the databases (cf. Section 5) to geometrically normalise the images. Images are then photometrically enhanced to reduce the influence of illumination, for example, using the method introduced in [11].

From these preprocessed images, features that are useful for face recognition are extracted. Over the last few decades, numerous algorithms have been developed to extract different kinds of features like eigenfaces [25], local binary patterns [26], scale-invariant feature transform (SIFT) features [27] and Gabor features [28]. In addition, the way to extract features from raw pixel values has also been studied [29]. Using these features, a recognition algorithm is then executed, for example, linear discriminant analysis [30], the Bayesian intra-personal/extra-personal classifier [31], support vector machines [32], elastic bunch graph matching [33] or local Gabor binary pattern histogram sequences [34]. In this work, we focus on a face recognition system that was one of the best performing systems in [35], which relies on an ISV modelling in a Gaussian mixture model (GMM) framework using discrete cosine transform (DCT) block features.

To ensure reproducibility and comparability of our face recognition system, we strictly follow the evaluation protocols defined by the MOBIO and SCface databases and solely use open source software [36, 35] to run our experiments. The database protocols define the setup of the face verification experiment by dividing the images into three groups: training set, development set and evaluation set. First, facial features are extracted from all images of the database. Next, the images from the training set are used to adapt the face recognition system to the conditions of the database. Then, for each client in the development set, the features of one or more of the client's images are used to enrol a client model. The features of the remaining images from the development set are used to probe the system by computing similarity scores between client models and probe features. Finally, the scores from the evaluation set are computed in a similar way. These scores can be directly used to compute the recognition performance of the system, but they can also be further processed by score normalisation, for example, ZT-norm or score calibration.

2.1 UBM-GMM modelling of DCT block features

As in [14], the features extracted from the preprocessed images are DCT block features. After the image is decomposed into several overlapping blocks, DCT features

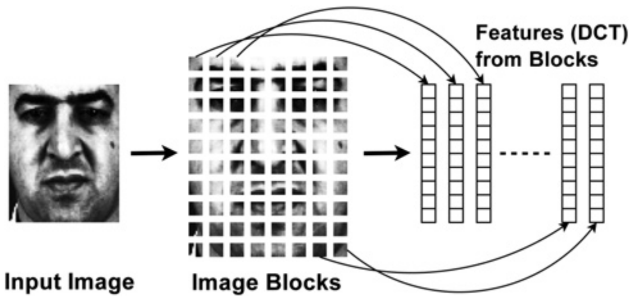

Fig. 1 *Process of extracting DCT block features from a geometrically normalised image*

$\boldsymbol{x}_b$ are extracted from each of the blocks. This extraction process is visualised in Fig. 1.

In contrast to most approaches to face recognition, these features are not concatenated into a single long feature vector, but each feature is taken to be an independent observation of the same person. To enrol a model of a client, the distribution of DCT block features from one or more images from the client is modelled by a GMM. The enrolment process to create the client-specific GMM is twofold. First, a client-unspecific GMM – the so-called universal background model (UBM) λ_{UBM} – models the distribution of features from an independent set of training images that does not include images from clients. Secondly, the client-specific GMM λ_c is created by adapting the means of the UBM to the features of the client's enrolment features [14] while keeping the same covariance matrices as the UBM.

2.2 ISV modelling

The ISV modelling technique was originally inspired by the speaker recognition field [37]. This technique involves estimating a linear subspace in GMM supervector space to capture the effects of image variations (due to, e.g. illumination, pose, facial expression and occlusion) and accounts for these variations during client model enrolment. The enrolled client-specific GMMs thereby isolate a client-specific component from image-dependent components in GMM supervector space. This modelling technique has been shown to improve stability against these image-dependent variations. For details, readers are directed to [14].

During the deployment (test) phase, the DCT features $\boldsymbol{x}_p = \left\{\boldsymbol{x}_{p,b}\right\}_{b=1}^{B}$ for all blocks b of a probe image are extracted, and an estimate is made of how well the probe features can be explained by a certain client model λ_c. Specifically, this is achieved by computing the average log-likelihood ratio (LLR) score

$$h\left(\boldsymbol{x}_p, \lambda_c\right) = \frac{1}{B}\sum_{b=1}^{B} \log \frac{p\left(\boldsymbol{x}_{p,b}|\lambda_c\right)}{p\left(\boldsymbol{x}_{p,b}|\lambda_{\text{UBM}}\right)} \tag{1}$$

This score, thus, compares the likelihood that the client model λ_c generated the observations (H_P) against the likelihood that they were generated by the UBM, λ_{UBM} (H_D).

2.3 ZT-score normalisation

After score computation, we employ ZT-norm, which was also adopted from the speaker verification field [38]. ZT-norm incorporates both client-centric Z-norm and probe-centric T-norm [39]. The goal of ZT-norm is to make the score independent of the current client or probe.

Both Z- and T-norm convert a raw score h to a normalised score h' by subtracting an average impostor score μ and dividing it by its standard deviation σ:

$$h' = \frac{h - \mu}{\sigma} \tag{2}$$

The difference between Z- and T-norm is how impostor scores are computed. For Z-norm, these scores are computed between the currently tested client model λ_c and all probe images from the cohort, whereas for T-norm, scores are computed between the current probe $\boldsymbol{x}_p$ and all cohort client models.

Finally, ZT-norm is a combination of first applying Z-norm and then applying T-norm afterwards, which was shown to perform well for face recognition [15]. It should be noted that the ZT-norm score transformation removes any LLR properties that the scores may have had before transformation.

3 Likelihood ratio calibration

Using an automatic face recognition system for forensic applications, it is important to ensure that scores are output in the form of LRs. Even if the face recognition algorithms are designed to produce LR scores, because of various reasons like score normalisation or imbalanced training data, this goal might not be directly achieved. One way to give LR properties to the face recognition scores is through calibration, which is described as 'the act of defining the mapping from score to LLR' [19].

3.1 Likelihood ratios for forensic face recognition

Experts argue that reporting an LR is a sound way of presenting scientific evidence to court. An LR expresses the ratio of two likelihoods. For forensics, this is the ratio of the likelihoods of observing the evidence E in two competing hypothesis: the prosecution hypothesis H_P and the defense hypothesis H_D

$$\text{LR} = \frac{P(E|H_P)}{P(E|H_D)} \tag{3}$$

For forensic face recognition, these two competing hypotheses can be defined as

- H_P: probe $\boldsymbol{x}_p$ originates from the client c, and
- H_D: probe $\boldsymbol{x}_p$ originates from someone else.

For numerical stability reasons, the LR is taken in the logarithmic domain, forming the LLR.

3.2 Linear score transformation

One way to perform calibration in a binary classification process like face verification is through linear calibration [40]. This calibration process linearly transforms raw scores produced by a face recognition system to calibrated LR scores. The linear transformation used to calibrate raw scores h (or h' after ZT-norm) to calibrated LLRs ℓ is

$$\ell = w_0 + w_1 h \tag{4}$$

IET Biom., 2014, Vol. 3, Iss. 4, pp. 246–256
doi: 10.1049/iet-bmt.2013.0066

where w_0 is the offset parameter and w_1 is the scaling parameter. These two parameters are obtained from the scores of the development set of the database via logistic regression.

Finally, the trained calibration parameters are applied to the scores of the evaluation set. In this way, calibration transfers knowledge about the whole score distribution from the development set to the evaluation set, in order to improve the interpretability of the resulting calibrated scores.

3.3 Categorical calibration

In this paper, we introduce a technique called categorical calibration to the face recognition field. This calibration technique is an extension of linear calibration described above that replaces the single offset parameter w_0 with a set of N category-dependent offset parameters $w_{0,i}$. Assuming that there are N distinct probe image categories $Q = \{q_i\}_{i=1}^{N}$ and that, therefore, probe features $\boldsymbol{x}_p$ that produced score h belong to a certain category q, scores transformation using categorical calibration can be formulated as

$$\ell = \sum_{i=1}^{N} \delta_{q,q_i} w_{0,i} + w_1 \, h \tag{5}$$

where δ is the Kronecker delta

$$\delta_{q,q_i} = \begin{cases} 1, & \text{if } q = q_i \\ 0, & \text{if } q \neq q_i \end{cases} \tag{6}$$

Categorical calibration is motivated by a calibration technique in speaker recognition that employs side information [41]. In categorical calibration, the categories can be in the form of quality measures [42, 43] of the image such as subject pose, illumination condition, resolution, facial expression, and so on. In this paper, we use distance between camera and subject to determine the category of probe images. Unlike conventional linear calibration, an improvement in verification performance is possible through categorical calibration. This is because the rank order of scores is invariant under (4) but not under (5).

4 Performance measures

Two types of metrics are used to measure the verification performance of our face recognition system. The metrics are verification cost (C_{ver}) and probability of false rejection (P_{fr}), both of which measure performance at different locations in the ROC curves, as well as the cost of LLR (C_{llr}), which assesses the whole ROC curve. In this section, we introduce these measures in more detail. For all metrics, lower values indicate better system performance.

4.1 Verification cost

The verification cost C_{ver} is a binary-classification system performance measure, which is defined as

$$\begin{aligned} C_{\text{ver}}(\theta) = {} & P_{\text{cli}} \times C_{\text{FR}} \times \text{FRR}(\theta) \\ & + (1 - P_{\text{cli}}) \times C_{\text{FA}} \times \text{FAR}(\theta) \end{aligned} \tag{7}$$

where P_{cli} is the prior probability that the probe image is of the client, C_{FR} and C_{FA} are the weighted cost of false reject and false alarm errors, respectively, and θ is the decision threshold of the system. This metric is analogous to detection cost (C_{det}) in the speaker recognition field [44]. It measures the verification cost at a single operating point of the DET-curve [45] or at a certain false rejection rate (FRR) or false acceptance rate (FAR) point.

If the prior probability $P_{\text{cli}} = 0.5$ and the same weighting cost for C_{FR} and C_{FA} are used ($C_{\text{FR}} = C_{\text{FA}} = 1$), (7) becomes

$$C_{\text{ver}}(\theta) = \frac{\text{FRR}(\theta) + \text{FAR}(\theta)}{2} \tag{8}$$

This function is identical to the half total error rate (HTER), which is a well-known evaluation measure commonly used in face recognition [15, 46]. In our experiments, we use two different ways to determine a threshold θ. First, the optimal threshold θ^* is computed based on the development and evaluation set independently, by minimising

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \; C_{\text{ver}}(\theta) \tag{9}$$

In this paper, we refer to the minimum verification cost as $C_{\text{ver}}^{\min} = C_{\text{ver}}(\theta^*)$.

To give a more realistic and unbiased evaluation of the verification cost on the evaluation set, we also compute the optimal threshold θ^* based on the development set and compute the C_{ver} of the evaluation set at that threshold. For brevity, we simply call this value C_{ver}.

In addition to the C_{ver} measure, we also report the FRR at the threshold, where the FAR = 1% as probability of false rejection (P_{fr}) for both development and evaluation set. Both $C_{\text{ver}}^{\min}$ and P_{fr} are solely discrimination performance measures that are insensitive to linear calibration.

4.2 Cost of LLR

The last performance measure used in this paper is the cost of LLR (C_{llr}). Unlike C_{ver} and P_{fr}, the C_{llr} is an application-independent verification measure [47]. Usually, in face and speaker verification systems, hard decisions are made by thresholding the scores. The C_{llr} includes the concept of expected cost and soft Bayes decision. This metric can be seen as an integral over all cost functions C_{ver} in (7) that is parameterised by P_{cli}, C_{FR} and C_{FA}, thereby assessing calibration at all thresholds θ.

The metric C_{llr} is a performance measure commonly used in speaker recognition, for example, in the NIST SRE plan [20]. It can be interpreted as a scalar measure that summarises the quality of the LR scores [48]. The C_{llr} is formulated as

$$\begin{aligned} C_{\text{llr}} = {} & \frac{1}{2N_{\text{cli}}} \sum_{h_i \in \{h_{\text{cli}}\}} \log_2 \left(1 + \exp(-h_i)\right) \\ & + \frac{1}{2N_{\text{imp}}} \sum_{h_j \in \{h_{\text{imp}}\}} \log_2 \left(1 + \exp(h_j)\right) \end{aligned} \tag{10}$$

where N_{cli} and N_{imp} are the number of client and impostor trials, respectively. The C_{llr} value can be expressed as the sum of a minimum C_{llr} value referred to as discrimination loss, $C_{\text{llr}}^{\min}$, plus calibration loss, C_{mc}

$$C_{\text{mc}} = C_{\text{llr}} - C_{\text{llr}}^{\min} \tag{11}$$

Discrimination loss $C_{\text{llr}}^{\min}$ and calibration loss C_{mc} indicate the verification and calibration performances of a system, respectively [47]. To compute a meaningful value of C_{llr}, it is

Table 1 Interpretations of C_{llr} values for system performance and LR scores [47]

\cllr\ value	System performance interpretation	Special LLR properties
0	Perfect verification system	LLR = $-\infty$ for impostors and LLR = ∞ for clients
0 < \cllr\ < 1	Well-calibrated system	$-\infty$ < LLR < ∞ and LLRs are well-calibrated
1	Reference verification system	LLR = 0 for impostors and clients
\cllr\ > 1	Badly calibrated system	No LLR interpretation possible

important that the scores are interpretable as LRs and, therefore, calibration is required before computing this measure.

The C_{llr} can also be seen as a validity measure of a biometric system, in that it indicates the quality and validity of the LRs produced by the system [49]. The interpretation of C_{llr} values are presented in Table 1. A perfect verification system has $C_{llr} = 0$, while a reference system has $C_{llr} = 1$. The perfect verification system always produces LLR = $-\infty$ for impostor scores and LLR = ∞ for client scores. In contrast, the reference system always produces LLR = 0, that is, it does not add any information in the forensic decision process. When a verification system has $C_{llr} > 1$, it is considered to be badly calibrated. The scores produced by this system are misleading if interpreted as LRs. If the calibration loss C_{mc} is removed from the C_{llr} value, we find the discrimination loss is $0 \le C_{llr}^{min} < 1$.

A well-calibrated system has $0 \le C_{llr} < 1$ and produces well-calibrated LRs. A well-calibrated LR ℓ has the interesting property that 'the likelihood ratio of the likelihood ratio is the likelihood ratio', which is referred to as idempotence [50, 51]

$$\ell = \log \frac{P(\ell|H_P)}{P(\ell|H_D)} \tag{12}$$

This explains that the log likelihood ratio of log likelihood ratio ℓ is the log likelihood ratio ℓ itself. One implication of (12) is that for $\ell = 0$, the likelihoods of both H_P and H_D are equal.

5 Databases and protocols

We evaluate face verification and calibration performance on two challenging image databases. Since we want to evaluate performance in forensic cases and there is no publicly available forensic database, we chose the MOBIO [52] and SCface [53] databases that contain images that are as close as possible to real forensic data. Samples of facial images from the databases are presented in Fig. 2. To have unbiased evaluations (see [54] for effects of biased evaluations), the clients of each database are divided into three different sets:

1. A training set: Images of this set are used to learn the parameters of the face recognition algorithm. Here, model training uses two-thirds of this training data, while the remaining one-third is used as cohort images and cohort clients for ZT-score normalisation. In total, we use 9600 and 688 facial images of 50 and 43 identities for the MOBIO and SCface, respectively.
2. A development set: These images are used to optimise meta-parameters of the algorithm. The scores obtained with this set are also used to train score calibration parameters.
3. An evaluation set: These images are used to compute the final verification and calibration performances.

5.1 MOBIO

The MOBIO database [52] is a multi-modal face and speech database containing video recordings from mobile devices. The database was collected in order to capture real-world scenarios for face and speaker authentication. In this paper, we use image data extracted from the database [http://www.idiap.ch/dataset/mobio].

The 150 clients of the MOBIO database are divided into training set (50), development set (42) and evaluation set (58 persons). The training set is further split into 34 clients that are used to train the face recognition system, and 16 persons in the ZT-norm cohort.

The database is accompanied by two protocols, which are based on gender: *male* and *female*. Client models are enrolled using features from five facial images per identity. Finally, client and impostor scores are computed by probing all client models with all probe images. The number of client and impostor trials are listed in Table 2. Owing to the low number of clients in the training set, the training of the face recognition system and the ZT-norm is always performed gender-independently. However, calibration is executed gender-dependently, following the gender-split as specified in the protocols.

5.2 SCface

The SCface database [53] represents an indoor monitoring scenario. The probe images were captured from different surveillance cameras with three subject-to-camera distances: 1 m (*close*), 2.6 m (*medium*) and 4.2 m (*far*). With about 10 pixels inter-eye-distance, the *far* condition has the lowest image resolution, while the *close* condition has a viewing

a b

Fig. 2 *Example images*

a MOBIO database
b SCface databases
In SCface the first image shows an enrolment sample, while remaining images are from the *close*, *medium* and *far* condition, respectively

IET Biom., 2014, Vol. 3, Iss. 4, pp. 246–256
doi: 10.1049/iet-bmt.2013.0066

Table 2 Number of client and impostor scores in MOBIO and SCface

Database	Protocol	(Client/impostor trials)	
		Development set	Evaluation set
MOBIO	male	(2520/57 960)	(3990/147 630)
	female	(1890/32 130)	(2100/39 900)
SCface	close	(220/9460)	(215/9030)
	medium	(220/9460)	(215/9030)
	far	(220/9460)	(215/9030)
	combined	(660/28 380)	(645/27 090)

angle slightly from above (cf. Fig. 2*b*). As is often the case in real surveillance applications, client models are each enrolled from a single high-quality frontal mug-shot photograph.

In total, the number of clients in the SCface database is 130. They are split into sets of 43 subjects for training, 44 for development and 43 for evaluation. The training clients are split up into 29 clients that are used to train the face recognition system and 14 identities in the cohort. There are four protocols defined: *close*, *medium*, *far* and *combined*. The *combined* protocol includes all images from the *close*, *medium* and *far* conditions. Again, all probe images are compared to all client models, leading to the number of client and impostor trials listed in Table 2.

6 Experimental setup

In this section, we describe the setup of the face recognition system and calibration. We execute experiments on both databases independently. For each database, the face recognition system is adapted to the training set of the database and the cohort images are taken only from the corresponding training set. The parameters for the face recognition experiments, explained in more detail in this section, are optimised to the development set of each database separately. Here we use the same algorithm configuration as in [15]. Except where stated otherwise, ZT score normalisation always uses cohort images across all conditions, that is, gender-independent for MOBIO and distance-independent for SCface.

Importantly, all results are generated solely using open source software. The face recognition algorithm, the linear calibration of scores, the verification and calibration metrics, as well as the image database interfaces rely on the open source signal-processing and machine learning toolbox Bob [36] [http://www.idiap.ch/software/bob]. The face recognition and linear calibration experiments are conducted with the FaceRecLib [35] [http://pypi.python.org/pypi/facereclib], which implements the evaluation protocols for the databases. The calibration module inside Bob is adapted from Bosaris [48], a toolkit for calibrating, fusing and evaluating scores from binary classifiers. All results, figures, tables and plots presented in this paper can be reproduced using the provided software package [http://pypi.python.org/pypi/xfacereclib.paper.IET2014].

6.1 Face recognition

The first step of the image processing chain for face recognition is image preprocessing. After geometrical alignment using the hand-labelled eye positions that are provided with the databases, the eye positions in the resulting grey-scale image are horizontally aligned at 16 pixels from the top and separated by 33 pixels, with a resulting image resolution of 64×80 pixels. To reduce the effects of illumination, the images of the MOBIO database are photometrically normalised [11].

The preprocessed images are split into overlapping blocks of 12×12 pixels for MOBIO and 20×20 pixels for SCface, sampled with the minimum step size of 1 pixel [15]. Thus, a total of $B = 3657$ or 2745 blocks are generated from each image in the MOBIO or SCface database, respectively.

Each image block is normalised such that pixel values have zero mean and unit variance. Then, from each image block a set of DCT features [55] is extracted, and the 45 (MOBIO) or 66 (SCface) lowest frequency components are retained. Finally, the coefficients of all blocks in every image are again normalised to zero mean and unit variance [15].

For the face recognition system, a separate UBM is computed for each of the two databases. To train the linear ISV subspace, we use the same training data as for UBM creation. As in [15], we selected a subspace of 320 dimensions for MOBIO and 80 dimensions for SCface.

6.2 Calibration

Two calibration conditions are evaluated in the MOBIO database. These conditions are based on gender division into *male* and *female* subsets. The calibration parameters are computed from the scores of the development set of each gender independently. Afterwards, calibration is applied to the scores of the evaluation set with corresponding gender.

Four distance conditions in the SCface database, which are *close*, *medium*, *far* and *combined*, are evaluated. Besides conventional linear calibration, we also apply categorical calibration to the *combined* scores of SCface. In this categorical calibration experiment, additional information about facial images, that is, the distance between surveillance camera and subject is used. Specifically, the distances *close*, *medium* and *far* are used to form the set of probe image categories Q.

7 Results

This section describes the results of our face recognition and score calibration experiments. Evaluated on the MOBIO and SCface databases, the verification performance of the face recognition system is observed with and without ZT-norm. Afterwards, calibration is applied to both raw and ZT-normalised scores. Categorical calibration is shown to be beneficial for both the discrimination and calibration performance of SCface scores. At the end of this section, we present a detailed analysis of the effect of calibration on score distributions.

7.1 Verification performance before calibration

The verification performance of the face recognition system for both the MOBIO and SCface databases is presented in Table 3. The performance is expressed in terms of $C_{\mathrm{ver}}^{\mathrm{min}}$ and P_{fr} for the development and evaluation set. Additionally, the unbiased C_{ver} measure is given for the evaluation set, where the optimal threshold θ^* from the development set is taken into account.

For the MOBIO database, the verification results for development and evaluation set differ. While in the development set the $C_{\mathrm{ver}}^{\mathrm{min}}$ values range around 4% for *male*

doi: 10.1049/iet-bmt.2013.0066

Table 3 Verification performance using raw and ZT-normalised scores, evaluated on MOBIO and SCface

Dataset (dev/eval)	Raw scores					ZT-norm				
	Dev. set		Eval. set			Dev. set		Eval. set		
	C_{ver}^{min}	P_{fr}	C_{ver}^{min}	C_{ver}	P_{fr}	C_{ver}^{min}	P_{fr}	C_{ver}^{min}	C_{ver}	P_{fr}
MOBIO:										
a. male	3.90%	**9.52%**	7.10%	7.26%	17.44%	**3.87%**	10.28%	**6.52%**	**6.77%**	**17.42%**
b. female	**5.84%**	**13.07%**	11.86%	**12.69%**	37.71%	6.87%	18.84%	**10.21%**	14.78%	**35.57%**
SCface:										
a. close	10.66%	30.91%	10.57%	10.82%	35.81%	**7.14%**	**27.27%**	**8.10%**	**8.74%**	**35.35%**
b. medium	11.19%	38.64%	8.08%	8.91%	33.02%	**9.32%**	**36.36%**	**6.90%**	**7.48%**	**32.56%**
c. far	19.39%	**73.64%**	19.99%	**20.45%**	**73.95%**	**18.40%**	74.55%	**19.66%**	20.51%	76.28%
d. combined	17.03%	52.27%	16.39%	16.41%	51.01%	**12.56%**	**45.15%**	**12.23%**	**12.44%**	**44.81%**

and 6% for *female* clients, they are 7% and 11%, respectively, in the evaluation set. This is similar to what has been observed in [15, 46]. ZT-norm improves the C_{ver}^{min} values for the evaluation set, but not for the development set of MOBIO *female* data. In this condition, there seems to be shift of scores from development to evaluation set, which causes relatively large differences between C_{ver}^{min} and C_{ver}. In addition, ZT-norm seems to maintain only the P_{fr} values.

For the SCface database, the four protocols *close*, *medium*, *far* and *combined* are evaluated. In Table 3, ZT-norm is performed using only cohort images from the corresponding distance condition. The *close* and *medium* images with sufficient image resolution provide C_{ver}^{min} error rates in the order of 10%, whereas in the *far* condition the error rates are roughly doubled. In general, ZT-norm improves the verification performance moderately, especially for the *combined* protocol where error rates are reduced by up to 4% after ZT-norm. This positive gain of ZT-norm can be observed across all performance measures in Table 3.

Motivated by the last observation, we repeated the ZT-norm experiments using cohort images across all distance conditions. The results of this experiment are shown in Table 4*a*. Interestingly, nearly all error rates dropped remarkably, except for the *far* condition, which seems to be little effected. Additionally, we tested how the selection of the threshold influences performance. In Table 3, the threshold is computed for each distance condition independently. In Table 4*b*, a single threshold for all conditions is selected. Clearly, the performance on the evaluation set drops seriously, especially for the *medium* and *close* conditions. The values of C_{ver}^{min} in Tables 3 and 4*b* are identical because the measure is independent of the threshold.

Table 4 Verification performance for the SCface database showing the impacts of (a) using all conditions for the ZT-norm cohort and (b) computing the threshold on the *combined* set without ZT-norm

Protocol	C_{ver}^{min}(dev)	C_{ver}^{min}(eval)	C_{ver}	P_{fr}
a. ZT-norm with *combined* cohort				
close	7.14%	8.10%	8.27%	29.77%
medium	9.32%	6.24%	6.61%	26.51%
far	18.40%	20.07%	20.78%	78.60%

Protocol	C_{ver}^{min}	C_{ver}
b. threshold on *combined* set		
close	10.57%	14.68%
medium	8.08%	13.75%
far	19.99%	20.79%

The observation from the last two experiments is that integrating additional information about the images, for example, the subject-to-camera distance into the face recognition system improves verification, but this is apparently not true for all the steps of the face recognition tool chain. Therefore, in the following calibration experiments, we use the best setup for the SCface database: ZT-norm uses cohort images across all distance conditions, whereas the threshold is based on distance-dependent scores.

7.2 Calibration performance

To study the effect of calibration on face recognition, the system performance is evaluated using the C_{llr} measure. The evaluated scores are the calibrated LRs from the evaluation sets of MOBIO and SCface. The C_{llr} measure is composed of the sum of two metrics: the discrimination loss C_{llr}^{min}, which reflects the minimum loss due to verification errors, and the calibration loss C_{mc}, which reflects the additional cost of miscalibration. The calibration experiment results are presented in Table 5. In general, C_{llr}^{min}values after ZT-norm are lower than those of raw scores, which indicates that better verification performance is offered by the ZT-norm scores. In the MOBIO database, for the ZT-norm scores there are 7 and 8% relative improvements in C_{llr}^{min} compared to raw scores for *male* and *female* genders, respectively. For SCface, the system with ZT-norm has improved C_{llr}^{min} discrimination performance compared to the raw system in most distance conditions. Stable performance is observed in *far* condition, while significant relative improvements are shown for other distance conditions, ranging from 17% in the *combined* condition to 40% for *close*. These observations are in line with the results reported in Section 7.1.

Table 5 Calibration performance after linear calibration of the raw and ZT-normalised scores of the evaluation set of MOBIO and SCface

Dataset condition: (eval. set)	Raw scores			ZT-norm		
	C_{llr}^{min}	C_{llr}	C_{mc}	C_{llr}^{min}	C_{llr}	C_{mc}
MOBIO:						
a. male	0.254	0.278	0.024	**0.236**	**0.257**	**0.021**
b. female	0.392	**0.473**	**0.080**	**0.360**	0.483	0.122
SCface:						
a. close	0.343	0.378	0.034	**0.261**	**0.287**	**0.026**
b. medium	0.284	0.313	**0.029**	**0.205**	**0.243**	0.038
c. far	**0.625**	**0.659**	0.034	0.636	0.664	**0.028**
d. combined	0.503	0.523	0.020	**0.419**	**0.432**	**0.013**

IET Biom., 2014, Vol. 3, Iss. 4, pp. 246–256
doi: 10.1049/iet-bmt.2013.0066

Table 6 C_{ver} (θ_0) values before and after calibration is applied to the ZT-normalised scores in the evaluation set of MOBIO and SCface

Dataset	C_{ver}^{min}	C_{ver}	C_{ver} (θ_0)	
			Before calibration	After calibration
MOBIO:				
a. male	6.52%	6.77%	35.93%	6.65%
b. female	10.21%	14.78%	38.08%	13.64%
SCface:				
a. close	8.10%	8.27%	26.37%	8.22%
b. medium	6.24%	6.61%	26.22%	6.42%
c. far	20.07%	20.78%	30.13%	20.62%
d. combined	12.23%	12.44%	27.57%	12.64%

Table 5 shows that ZT-normalisation improves C_{llr} compared to raw scores, except for the *female* condition in MOBIO. Apparently, applying ZT-norm results in an improved C_{llr}^{min}, but not necessarily an improved C_{mc}. This means that applying ZT-norm reduces discrimination loss, while the effect of calibration loss (C_{mc}) results in an inferior C_{llr} for the *female* subset of MOBIO compared to the raw scores.

Table 6 presents the verification cost C_{ver} at threshold $\theta_0 = 0$, which is computed before and after calibration for the ZT-normalised scores from the evaluation set of MOBIO and SCface. Threshold $\theta_0 = 0$ is selected as it represents the application-independent threshold for well-calibrated LR scores. In Table 6, it is clearly shown that the $C_{ver}(\theta_0)$ values after calibration are far lower than before calibration. Mostly, $C_{ver}(\theta_0)$ values are in the order of the C_{ver} values or even lower, which shows that calibration can produce well-calibrated LRs from the ZT-normalised scores that are produced by our face recognition system.

From our evaluation using C_{llr}, it has been found that ZT-norm is favoured to increase face recognition performance in general. Through calibration, raw scores from the face recognition system have been successfully converted into LLR scores so that $\theta_0 = 0$ becomes a valid threshold as measured by the verification performance metric C_{ver}.

7.3 Categorical calibration in SCface

In the experiment with categorical calibration, we include the distance information of SCface images as categories $Q = \{close, medium, far\}$ to improve calibration and verification performance of the face recognition system. For categorical calibration, the scores from the *combined* distance condition with ZT-score normalisation are used.

The results of this categorical experiment are presented in Table 7. In the first row, the values of C_{llr} and C_{mc} are presented for uncalibrated scores for the sake of completeness. The reader should bear in mind that the metric C_{llr} is only meaningful for evaluating scores with an LR interpretation.

Comparing the performance of linear and categorical calibration, the latter provides a relative reduction in C_{llr}^{min} and C_{llr} of around 6%. In general, including category information through categorical calibration improves verification performance. Based on the C_{ver} values in Table 7, categorical calibration has successfully improved verification performance compared to linear calibration, by 5.2% in C_{ver}^{min} and 2.7% in C_{ver}. Similarly, categorical calibration performs well in terms of $C_{ver}(\theta_0)$ with relative improvement of 6.4%. In terms of P_{fr}, however, the categorical calibration can only maintain the system verification performance. This effect might be explained by the fact that categorical calibration focuses on the overlapping part of the score distributions, and not on the tail belonging low FAR values.

The findings in this categorical calibration experiment show that the categorical calibration technique, in general, offers better face recognition performance in both verification and calibration compared to the linear calibration technique.

7.4 Discussion

In the previous sections, we analysed the verification and calibration performance of the face recognition system with regards to the use of ZT-norm. It was shown that ZT-norm, in general, helps to improve the verification performance. Furthermore, both linear and categorical calibration were applied to the scores, resulting in improved calibration performance. In this section, we further analyse the effect of calibration with respect to the distribution of client and impostor scores.

The score distributions for the evaluation set of both MOBIO and SCface before and after calibration are presented in Fig. 3. The distributions are depicted for the *male* gender in MOBIO and the *combined* distance condition in SCface. ZT-norm affects distribution of uncalibrated scores for both MOBIO and SCface (first column of Fig. 3). Generally, both raw and ZT-normalised impostor scores assemble around score value 0 before calibration. For SCface, the raw scores show a high peak compared to the ZT-normalised uncalibrated scores.

Depicted in the second column of Fig. 3, the distributions of calibrated LLR scores represent the behaviour of well-calibrated LLRs. One indicator is the intersection between the score distribution of clients and impostors, which lies near the LR $\ell = 0$. This corresponds to the properties of well-calibrated LLR ℓ explained in (12).

In addition to the analysis of score distributions before and after linear calibration, we present the score distributions after categorical calibration. In Fig. 4, the score distributions for the SCface evaluation set with ZT-norm are depicted before calibration, after linear calibration and after categorical calibration. Both linear and categorical calibration scale and shift the score distributions such that the intersection of the

Table 7 Verification and calibration performance of the ZT-normalised scores of the SCface *combined* protocol before calibration and after linear and categorical calibration

Calibration technique	C_{llr}^{min}	C_{llr}	C_{mc}	C_{ver}^{min}, %	C_{ver}, %	C_{ver} (θ_0), %	P_{fr}, %	# param
none	0.419	0.736	0.317	12.23	12.44	27.57	**44.81**	0
linear	0.419	0.432	**0.013**	12.23	12.44	12.64	**44.81**	2
categorical	**0.392**	**0.406**	0.014	**11.59**	**12.11**	**11.83**	47.13	5

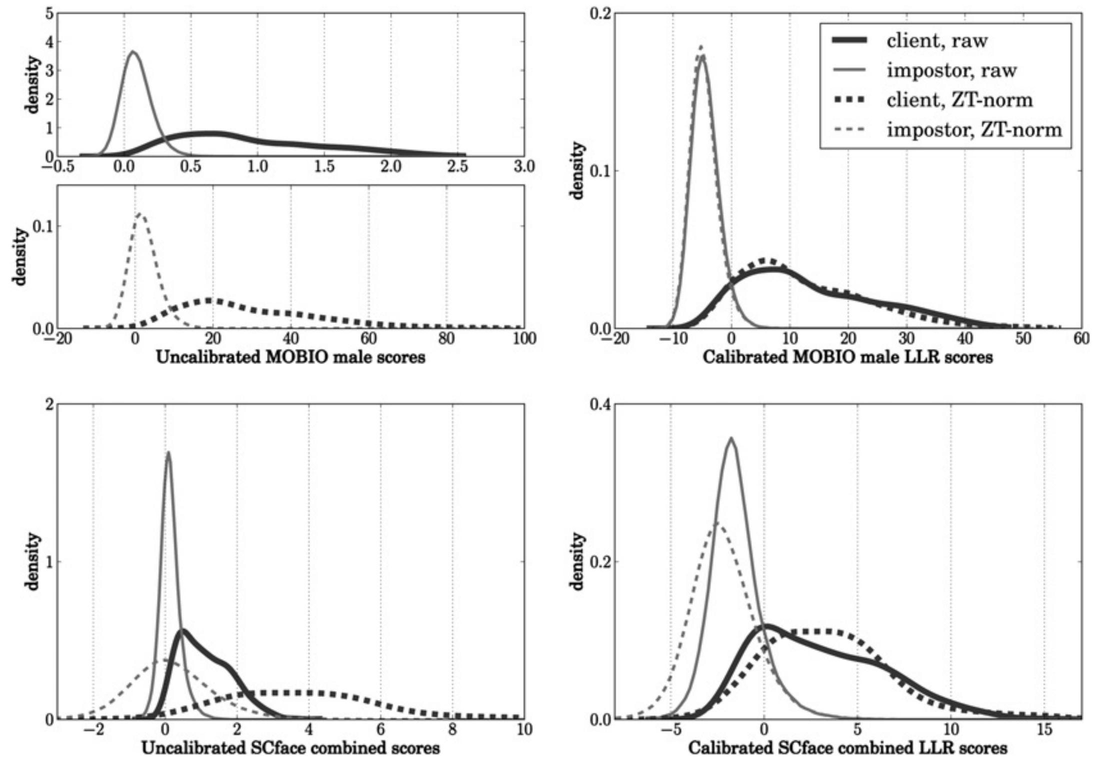

Fig. 3 *Score distributions for MOBIO male and SCface combined before and after calibration, both before (raw) and after ZT score normalisation (ZT-norm)*

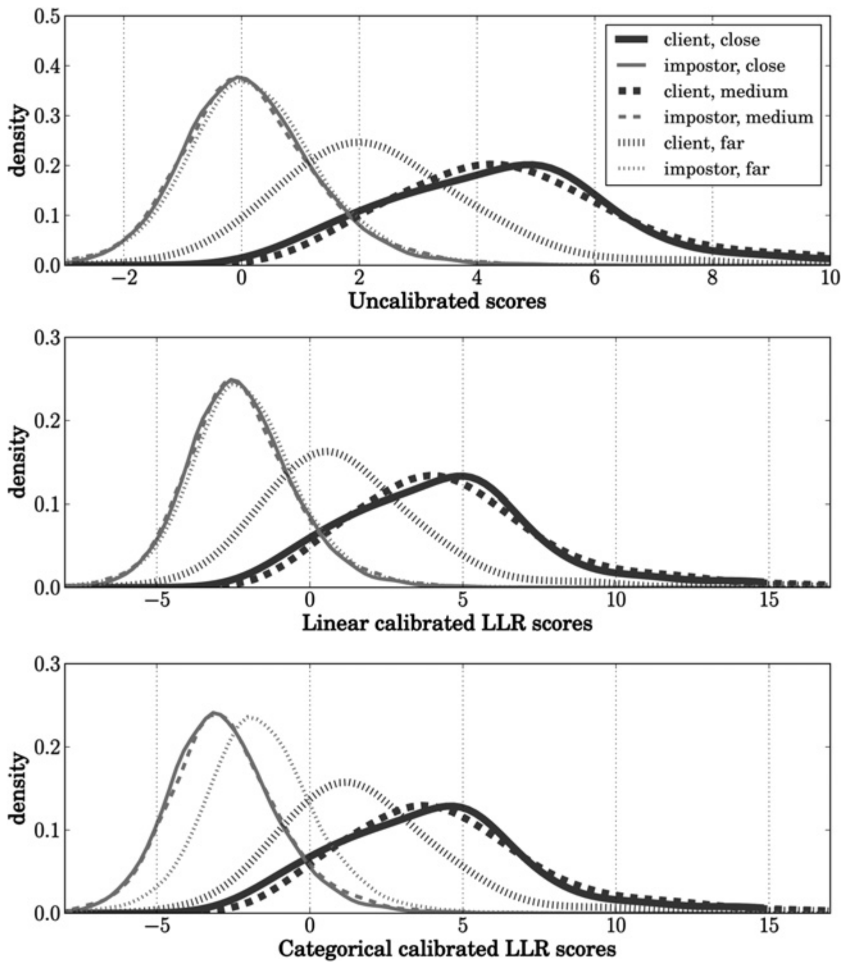

Fig. 4 *Distributions of scores from SCface with ZT-normalisation before calibration, after linear calibration and after categorical calibration*

doi: 10.1049/iet-bmt.2013.0066

client and impostor distributions lies closer to $\ell = 0$. Especially for the categorical calibration, all three different distance conditions intersect exactly at $\ell = 0$. This shows that both calibration techniques have successfully produced well-calibrated LRs from the ZT-normalised scores. As described previously, a common scaling parameter w_1 is utilised in (5) for all categories, *close*, *medium* and *far*, while a different offset $w_{0,i}$ is used for each category. Fig. 4 illustrates how this extra information and flexibility in calibration results in improved separation and distribution of scores, ultimately leading to improved verification and calibration performance.

8 Conclusion

In this paper, we presented evaluations of calibration of a face recognition system based on ISV modelling on the MOBIO and SCface databases. Calibration produces scores in the form of LRs. We performed categorical calibration on the SCface database with subject-to-camera distance as a category. We showed that categorical calibration improves face recognition performance in terms of calibration and verification compared to a system with linear calibration, by incorporating additional information about the probe images in the calibration process.

Through this paper, we hope to encourage further research in the area of calibration for face recognition using the categorical calibration technique, since it can be applied to other categories such as pose, illumination and expression to reduce the impact of these image variations from the face recognition process. Researchers are encouraged to utilise our open source software package, which is easily understandable, well-documented and tested.

9 Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 238803.

10 References

1 Bolle, R.M., Connell, J.H., Pankanti, S., Ratha, N.K., Senior, A.W.: 'The common biometrics: Guide to biometrics' (Springer, 2004)
2 Peacock, C., Goode, A., Brett, A.: 'Automatic forensic face recognition from digital images', *Sci. Justice: J. Forensic Sci. Soc.*, 2004, **44**, (1), pp. 29
3 Jain, A.K., Ross, A., Prabhakar, S.: 'An introduction to biometric recognition', *IEEE Trans. Circuits Syst. Video Technol.*, 2004, **14**, (1), pp. 4–20
4 Wright, D.B., Memon, A., Skagerberg, E.M., Gabbert, F.: 'When eyewitnesses talk', *Curr. Dir. Psychol. Sci.*, 2009, **18**, (3), pp. 174–178
5 Loftus, E.F., Hoffman, H.G.: 'Misinformation and memory: the creation of new memories', *J. Exp. Psychol. Gen.*, 1989, **118**, (1), pp. 100
6 Loftus, E.F., Miller, D.G., Burns, H.J.: 'Semantic integration of verbal information into a visual memory', *J. Exp. Psychol. Hum. Learn. Mem.*, 1978, **4**, (1), p. 19
7 Ellis, H.D., Shepherd, J.W., Davies, G.M.: 'Identification of familiar and unfamiliar faces from internal and external features: some implications for theories of face recognition', *Perception*, 1979, **8**, (4), pp. 431–439
8 Burton, A.M., Wilson, S., Cowan, M., Bruce, V.: 'Face recognition in poor-quality video: evidence from security surveillance', *Psychol. Sci.*, 1999, **10**, (3), pp. 243–248
9 O'Toole, A.J., Phillips, P.J., Jiang, F., Ayyad, J., Penard, N., Abdi, H.: 'Face recognition algorithms surpass humans matching faces over changes in illumination', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2007, **29**, (9), pp. 1642–1646
10 Jafri, R., Arabnia, H.R.: 'A survey of face recognition techniques', *JIPS*, 2009, **5**, (2), pp. 41–68
11 Tan, X., Triggs, B.: 'Enhanced local texture feature sets for face recognition under difficult lighting conditions', *Anal. Model. Faces Gestures*, 2007, pp. 168–182
12 Viikki, O., Laurila, K.: 'Cepstral domain segmental feature vector normalization for noise robust speech recognition', *Speech Commun.*, 1998, **25**, (1), pp. 133–147
13 Pelecanos, J., Sridharan, S.: 'Feature warping for robust speaker verification'. Odyssey: The Speaker and Language Recognition Workshop, International Speech Communication Association (ISCA), 2001, pp. 213–218
14 Wallace, R., McLaren, M., McCool, C., Marcel, S.: 'Inter-session variability modelling and joint factor analysis for face authentication'. Int. Joint Conf. Biometrics (IJCB), 2011, pp. 1–8
15 Wallace, R., McLaren, M., McCool, C., Marcel, S.: 'Cross-pollination of normalization techniques from speaker to face authentication using Gaussian mixture models', *IEEE Trans. Inf. Forensics Sec.*, 2012, **7**, (2), pp. 553–562
16 Champod, C., Meuwly, D.: 'The inference of identity in forensic speaker recognition', *Speech Commun.*, 2000, **31**, (2), pp. 193–203
17 Ramos-Castro, D., Gonzalez-Rodriguez, J., Ortega-Garcia, J.: 'Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework'. Odyssey: The Speaker and Language Recognition Workshop, IEEE, International Speech Communication Association (ISCA), 2006, pp. 1–8
18 Castro, D.R.: 'Forensic evaluation of the evidence using automatic speaker recognition systems'. PhD thesis, Universidad autónoma de Madrid, 2007
19 Brümmer, N., du Preez, J.: 'Application-independent evaluation of speaker detection', *Comput. Speech Lang.*, 2006, **20**, (2), pp. 230–275
20 National Institute of Standards and Technology: The NIST Year 2010 Speaker Recognition Evaluation Plan. Available at: http://www.nist.gov/itl/iad/mig/sre12.cfm
21 Gonzalez-Rodriguez, J., Fierrez-Aguilar, J., Ramos-Castro, D., Ortega-Garcia, J.: 'Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems', *Forensic Sci. Int.*, 2005, **155**, (2), pp. 126–140
22 Champod, I.C., Evett, I.W., Kuchler, B.: 'Earmarks as evidence: a critical review', *J. Forensic Sci.*, 2001, **46**, (6), pp. 1275
23 Champod, C., Evett, I.W.: 'A probabilistic approach to fingerprint evidence', *J. Forensic Identif.*, 2001, **51**, (2), pp. 101–122
24 Poh, N., Tistarelli, M.: 'Customizing biometric authentication systems via discriminative score calibration'. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2681–2686
25 Turk, M., Pentland, A.: 'Eigenfaces for recognition', *J. Cognit. Neurosci.*, 1991, **3**, (1), pp. 71–86
26 Ahonen, T., Hadid, A., Pietikainen, M.: 'Face recognition with local binary patterns'. European Conf. Computer Vision, Proc. Workshop on Dynamical Vision, 2004, pp. 469–481
27 Križaj, J., Štruc, V., Pavešič, N.: 'Adaptation of SIFT features for robust face recognition'. ICIAR10, 2010, pp. 394–404
28 Daugman, J.G.: 'Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters', *J. Opt. Soc. Am. A*, 1985, **2**, (7), pp. 1160–1169
29 Cox, D.D., Pinto, N.: 'Beyond simple features: a large-scale feature search approach to unconstrained face recognition', 2011
30 Zhao, W., Krishnaswamy, A., Chellappa, R., Swets, D.L., Weng, J.: 'Discriminant analysis of principal components for face recognition' (Springer, Berlin, 1998), pp. 73–85. http://www.face-rec.org/algorithms/LDA/zhao98discriminant.pdf
31 Moghaddam, B., Wahid, W., Pentland, A.: 'Beyond eigenfaces: probabilistic matching for face recognition'. IEEE Int. Conf. Automatic Face and Gesture Recognition, 1998, pp. 30–35
32 Phillips, P.J.: 'Support vector machines applied to face recognition'. Advances in Neural Information Processing Systems (MIT Press, 1999), vol. 11, pp. 803–809
33 Wiskott, L., Fellous, J.-M., Krüger, N., Malsburg, C.v.d.: 'Face recognition by elastic bunch graph matching', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1997, **19**, pp. 775–779
34 Zhang, W., Shan, S., Gao, W., Chen, X., Zhang, H.: 'Local Gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition'. IEEE Int. Conf. Computer Vision, 2005, vol. 1, pp. 786–791
35 Günther, M., Wallace, R., Marcel, S.: 'An open source framework for standardized comparisons of face recognition algorithms', in: Fusiello, A., Murino, V., Cucchiara, R. (eds.): Computer Vision – ECCV 2012. Workshops and Demonstrations, Volume 7585 of Lecture Notes in Computer Science, Berlin, October 2012, pp. 547–556
36 Anjos, A., El Shafey, L., Wallace, R., Günther, M., McCool, C., Marcel, S.: 'Bob: a free signal processing and machine learning toolbox for

researchers'. 20th ACM Conf. Multimedia Systems (ACMMM), Nara, Japan, October 2012

37 Vogt, R., Sridharan, S.: 'Explicit modelling of session variability for speaker verification', *Comput. Speech Lang.*, 2008, **22**, (1), pp. 17–38

38 Zheng, R., Zhang, S., Xu, B.: 'A comparative study of feature and score normalization for speaker verification'. Proc. 2006 Int. Conf. Advances in Biometrics, ICB'06, Berlin, Heidelberg, 2006, pp. 531–538

39 Auckenthaler, R., Carey, M., Lloyd-Thomas, H.: 'Score normalization for text-independent speaker verification systems', *Digit. Signal Process.*, 2000, **10**, (1), pp. 42–54

40 Brümmer, N., Burget, L., Cernocky, J.H., *et al.*: 'Fusion of heterogeneous speaker recognition systems in the stub submission for the NIST speaker recognition evaluation 2006', *IEEE Trans. Audio Speech Lang. Process.*, 2007, **15**, (7), pp. 2072–2084

41 van Leeuwen, D.A.: 'The TNO SRE-2008 speaker recognition system'. Proc. NIST Speaker Recognition Evaluation Workshop, Montreal, 2009

42 Garcia-Romero, D., Fierrez-Aguilar, J., Gonzalez-Rodriguez, J., Ortega-Garcia, J.: 'On the use of quality measures for text-independent speaker recognition'. Odyssey: The Speaker and Language Recognition Workshop. International Speech Communication Association (ISCA), 2004

43 Garcia-Romero, D., Fierrez-Aguilar, J., Gonzalez-Rodriguez, J., Ortega-Garcia, J.: 'Using quality measures for multilevel speaker recognition', *Comput. Speech Lang.*, 2006, **20**, (2), pp. 192–209

44 Doddington, G.R., Przybocki, M.A., Martin, A.F., Reynolds, D.A.: 'The NIST speaker recognition evaluation – overview, methodology, systems, results, perspective', *Speech Commun.*, 2000, **31**, (2), pp. 225–254

45 Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M.: 'The DET curve in assessment of detection task performance'. Technical Report, DTIC Document, 1997

46 Günther, M., Costa-Pazo, A., Ding, C., *et al.*: 'The 2013 face recognition evaluation in mobile environment'. The Sixth IAPR Int. Conf. Biometrics, 2013

47 van Leeuwen, D.A., Brümmer, N.: 'An introduction to application-independent evaluation of speaker recognition systems'. Speaker classification I (Springer, 2007), pp. 330–353

48 Brümmer, N., de Villiers, E.: 'The Bosaris toolkit: theory, algorithms and code for surviving the new DCF'. arXiv preprint arXiv:1304.2865, 2013

49 Morrison, G.S.: 'Measuring the validity and reliability of forensic likelihood-ratio systems', *Sci. Justice*, 2011, **51**, (3), pp. 91–98

50 van Leeuwen, D.A., Brümmer, N.: 'The distribution of calibrated likelihood-ratios in speaker recognition', 2013, Interspeech

51 Mandasari, M.I., Saeidi, R., McLaren, M., van Leeuwen, D.A.: 'Quality measure functions for calibration of speaker recognition system in various duration conditions', *IEEE Trans. Audio Speech Lang. Process.*, 2013

52 McCool, C., Marcel, S., Hadid, A., *et al.*: 'Bi-modal person recognition on a mobile phone: using mobile phone data'. IEEE ICME Workshop on Hot Topics in Mobile Multimedia, July 2012, pp. 635–640

53 Grgic, M., Delac, K., Grgic, S.: 'SCface – surveillance cameras face database', *Multim. Tools Appl.*, 2011, **51**, (3), pp. 863–879

54 Lui, Y.M., Bolme, D.S., Phillips, P.J., Beveridge, J.R., Draper, B.A.: 'Preliminary studies on the good, the bad, and the ugly face recognition challenge problem'. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2012, pp. 9–16

55 Cardinaux, F., Sanderson, C., Marcel, S.: 'Comparison of MLP and GMM classifiers for face verification on XM2VTS'. Fourth Int. Conf. Audio- and Video-Based Biometric Person Authentication, University of Surrey, Guildford, UK, 2003

doi: 10.1049/iet-bmt.2013.0066