Information Fusion 83–84 (2022) 110–130

Contents lists available at ScienceDirect

Information Fusion

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

Cross-sensor periocular biometrics in a global pandemic: Comparative benchmark and novel multialgorithmic approach

Fernando Alonso-Fernandez [a,*], Kiran B. Raja [b], R. Raghavendra [b], Christoph Busch [b], Josef Bigun [a], Ruben Vera-Rodriguez [c], Julian Fierrez [c]

[a] School of Information Technology, Halmstad University, Sweden
[b] Norwegian University of Science and Technology, Gjøvik, Norway
[c] School of Engineering, Universidad Autonoma de Madrid, Spain

ARTICLE INFO

Keywords:
Periocular recognition
Sensor interoperability
Cross-spectral
Cross-sensor
Ocular biometrics
Multibiometrics fusion
Linear logistic regression

ABSTRACT

The massive availability of cameras and personal devices results in a wide variability between imaging conditions, producing large intra-class variations and a significant performance drop if images from heterogeneous environments are compared for person recognition purposes. However, as biometric solutions are extensively deployed, it will be common to replace acquisition hardware as it is damaged or newer designs appear or to exchange information between agencies or applications operating in different environments. Furthermore, variations in imaging spectral bands can also occur. For example, face images are typically acquired in the visible (VIS) spectrum, while iris images are usually captured in the near-infrared (NIR) spectrum. However, cross-spectrum comparison may be needed if, for example, a face image obtained from a surveillance camera needs to be compared against a legacy database of iris imagery. Here, we propose a multialgorithmic approach to cope with periocular images captured with different sensors. With face masks in the front line to fight against the COVID-19 pandemic, periocular recognition is regaining popularity since it is the only region of the face that remains visible. As a solution to the mentioned cross-sensor issues, we integrate different biometric comparators using a score fusion scheme based on linear logistic regression This approach is trained to improve the discriminating ability and, at the same time, to encourage that fused scores are represented by log-likelihood ratios. This allows easy interpretation of output scores and the use of Bayes thresholds for optimal decision-making since scores from different comparators are in the same probabilistic range. We evaluate our approach in the context of the 1st Cross-Spectral Iris/Periocular Competition, whose aim was to compare person recognition approaches when periocular data from visible and near-infrared images is matched. The proposed fusion approach achieves reductions in the error rates of up to 30%–40% in cross-spectral NIR–VIS comparisons with respect to the best individual system, leading to an EER of 0.2% and a FRR of just 0.47% at FAR = 0.01%. It also represents the best overall approach of the mentioned competition. Experiments are also reported with a database of VIS images from two different smartphones as well, achieving even bigger relative improvements and similar performance numbers. We also discuss the proposed approach from the point of view of template size and computation times, with the most computationally heavy comparator playing an important role in the results. Lastly, the proposed method is shown to outperform other popular fusion approaches in multibiometrics, such as the average of scores, Support Vector Machines, or Random Forest.

1. Introduction

Periocular biometrics has gained attention during the last years as an independent modality for person recognition [1,2] after concerns of the performance of face or iris modality under non-ideal or uncooperative conditions [3,4]. The mandatory use of face masks due to the COVID-19 pandemic has produced that, even in cooperative settings, face recognition systems are presented with occluded faces where the periocular region is often the only visible area. This face occlusion comes with a reduction in facial information that may be significant for recognition [5,6]. To what extent this information reduction is detrimental for face recognition is yet something largely unexplored. In practice, recent studies have shown that commercial face recognition

* Corresponding author.
E-mail addresses: feralo@hh.se (F. Alonso-Fernandez), kiran.raja@ntnu.no (K.B. Raja), raghavendra.ramachandra@ntnu.no (R. Raghavendra), christoph.busch@ntnu.no (C. Busch), josef.bigun@hh.se (J. Bigun), ruben.vera@uam.es (R. Vera-Rodriguez), julian.fierrez@uam.es (J. Fierrez).

https://doi.org/10.1016/j.inffus.2022.03.008
Received 28 October 2020; Received in revised form 8 December 2021; Accepted 18 March 2022
Available online 29 March 2022

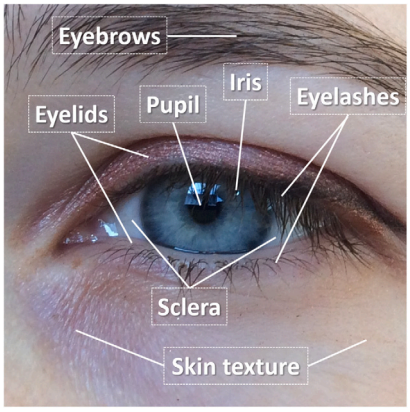

Fig. 1. Eye image labelled with some parts of the ocular region.

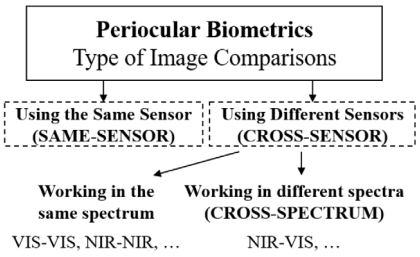

Fig. 2. Sensor interoperability in periocular biometrics.

engines, even in cooperative settings, struggle with persons wearing face masks [7], driving vendors to include capabilities for recognition of masked faces in their products [8]. In parallel, hygiene concerns are triggering fears against the use of contact-based biometric solutions such as fingerprints [9].

According to the Merriam-Webster dictionary, the medical definition of "periocular" is "surrounding the eyeball but within the orbit". From a forensic/biometric application perspective, our goal is to improve the recognition performance by using information extracted from the face region in the immediate vicinity of the eye, including the sclera, eyelids, eyelashes, eyebrows and the surrounding skin (Fig. 1). This information may include textural descriptors, but also the shape of the eyebrows or eyelids, or colour information [1]. With a surprising high discrimination ability, the resulting modality is the ocular one requiring the least constrained acquisition. It is sufficiently visible over a wide range of distances, even under partial face occlusion (close distance) or low-resolution iris (long distance), facilitating increased performance in unconstrained or uncooperative scenarios. It also avoids the need for iris segmentation, an issue in difficult images [10]. The COVID-19 outbreak has imposed the necessity of dealing with partially occluded faces even in cooperative applications in security, healthcare, border control or education. Another advantage in the context of the current global pandemic is that the periocular region appears in iris and face images, so it can be easily obtained with existing setups for face and iris.

The ocular region consists of several organs such as the cornea, pupil, iris, sclera, lens, retina, optical nerve, and periocular region. Some of them are shown in Fig. 1. Among these, iris, sclera, retina and periocular have been studied as biometric modalities [2]. The significant progress of ocular biometrics in the last decade has been primarily due to efforts in iris recognition since the late 80 s, resulting in large-scale deployments [11]. Iris provides very high accuracy with near-infrared (NIR) illumination and controlled, close-up acquisition. However, deployment to non-controlled environments is not yet mature due to the impact of low resolution, variable illumination, or off-angle views, which makes very difficult to locate and segment the iris [10]. Even if the latter can be achieved, the quality of the resulting iris image might not be sufficient for accurate recognition either [12]. The feasibility of vasculature of the sclera as a biometric modality (sometimes simply referred to as sclera) has also been established by several studies [13], although its acquisition in non-controlled environments poses the same problems as the iris modality. The vasculature of the retina is also very discriminative, and the retina is regarded as the most secure biometric modality due to being extremely difficult to spoof. However, its acquisition is very invasive, requiring high user cooperation and specialized optical devices.

In this context, periocular has rapidly evolved as a very popular modality for unconstrained biometrics [1,2,13], and recently due to the use of face masks even in constrained settings [7]. The term periocular is used loosely in the literature to refer to the externally visible region of the face that surrounds the eye socket. Therefore, images of the whole eye, such as the one in Fig. 1, are employed as input [13]. While the iris, sclera and other elements are present in such images, they are not explicitly used in isolation. It may be that the iris texture or the vasculature of the sclera cannot be reliably obtained either to be used as stand-alone modalities [12]. Some works even suggest that with visible light data, recognition performance is improved if components inside the ocular globe (iris and sclera) are discarded [14]. The fast-growing uptake of face technologies in social networks and smartphones, as well as the widespread use of surveillance cameras or face masks, has arguably increased the interest in periocular biometrics, especially in the visible (VIS) range. In such scenarios, samples captured with different sensors are to be compared if, for example, users are allowed to use their own acquisition devices, leading to a *cross-sensor* comparison in the same spectrum (VIS–VIS in this case). Unfortunately, this massive availability of cameras results in heterogeneous quality between images [15], which is known to decrease recognition performance significantly [11]. These sensor *interoperability* issues also arise when a biometric sensor is replaced with a newer one without reacquiring the corresponding template, thus forcing biometric samples from different sensors to co-exist. Sensors may also operate in a range other than VIS, such as NIR, leading to cross-sensor NIR–NIR comparisons, e.g. [16]. In addition, iris images are largely acquired beyond the visible spectrum [17], mainly using NIR illumination, but there are several scenarios in which it may be necessary to compare them with periocular images in the VIS range, leading in this case to a *cross-sensor* comparison in different spectra (NIR–VIS in this case), also known as *cross-spectral* comparison. This happens, for example, in law enforcement scenarios where the only available image of a suspect is obtained with a surveillance camera in the VIS range, but the reference database contains images in the NIR range [18,19]. These interoperability problems, if not properly addressed, can affect the recognition performance dramatically. Unfortunately, widespread deployment of biometric technologies will inevitably cause the replacement of hardware parts as they are damaged, or newer designs appear. Another application case is the exchange of information among agencies or applications which employ different technological solutions or whose data is captured in heterogeneous environments. The different types of image comparisons mentioned, based on the spectrum in which they have been acquired, are summarized in Fig. 2.

Accordingly, to counteract the reduction in recognition performance that is usually observed when comparing data from different sensors, we propose to combine the output of different periocular comparators at the score level, referred to as multialgorithm fusion (in contrast to multimodal fusion, which combines information from different modalities) [20,21]. The consolidation of identity evidence from heterogeneous comparators (also called experts, feature extraction techniques, or systems in the present paper) is known to increase recognition performance, because the different sources can compensate for the limitations of the others [20,22]. Integration at the score level is the most common approach because it only needs the output scores of

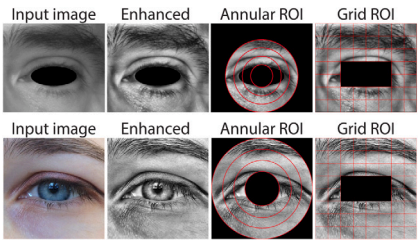

Fig. 3. Example images from Cross-Eyed (top row) and VSSIRIS (bottom row) databases. First column: input image. Second: after applying CLAHE (see Section 5.1). Third and fourth: ROI of the different biometric comparators (see Section 3).

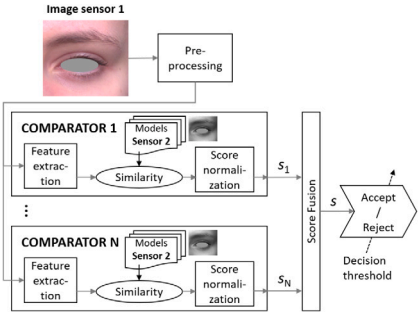

Fig. 4. Architecture of the proposed fusion strategy.

the different comparators, greatly facilitating the integration. With this motivation, we employ a multialgorithm fusion approach to cope with periocular images from different sensors which integrates scores from different comparators. It follows a probabilistic fusion approach based on linear logistic regression [23], in which the output scores of multiple systems are combined to produce a log-likelihood ratio according to a probabilistic Bayesian framework. This allows easy interpretation of output scores and the use of Bayes thresholds for optimal decision-making. This fusion scheme is compared with a set of simple and trained fusion rules widely employed in multibiometrics based on the arithmetic average of normalized scores [24], Support Vector Machines [25], and Random Forest [26].

The fusion approach based on linear logistic regression served as an inspiration to our submission to the 1st Cross-Spectral Iris/Periocular Competition (Cross-Eyed 2016) [27], with an outstanding recognition accuracy: Equal Error Rate (EER) of 0.29%, and False Rejection Rate (FRR) of 0% at a False Acceptance Rate (FAR) of 0.01%, resulting in the best overall competing submission. This competition was aimed at evaluating the capability of periocular recognition algorithms to compare visible and near-infrared images (NIR–VIS). In the present paper, we also carry out cross-sensor experiments with periocular images in the visible range only (VIS–VIS), but with two different sensors. For this purpose, we employ a database captured with two smartphones [28], demonstrating the benefits of the proposed approach to smartphone-based biometrics as well.

The rest of the paper is organized as follows. This introduction is completed with a description of the paper contributions. A summary of related works in periocular biometrics is given in Section 2. Section 3 then describes the periocular comparators employed. The score fusion methods evaluated are described in Section 4. Recognition experiments using images in different spectra (cross-spectral NIR–VIS) and in the visible spectrum (cross-sensor VIS–VIS) are described in Sections 5 and 6, respectively, including the databases, protocol used, results of the individual comparators, and fusion experiments. Finally, conclusions are given in Section 7.

1.1. Contributions

The contribution of this paper to the state of the art is thus as follows. First, we summarize related works in periocular biometrics using images from different sensors. Second, we evaluate nine periocular recognition comparators under the frameworks of different spectra (NIR–VIS) and same spectrum (VIS–VIS) recognition. The Reading Cross-Spectral Iris/Periocular Dataset (Cross-Eyed) [27] and the Visible Spectrum Smartphone Iris (VSSIRIS) [28] databases are respectively used for this purpose. We employ the three most widely used comparators in periocular research, which are used as a baseline in many studies [1]: Histogram of Oriented Gradients (HOG) [29], Local Binary Patterns (LBP) [30], and Scale-Invariant Feature Transform (SIFT) key-points [31]. Three other periocular comparators, proposed and published previously by the authors, are based on Symmetry Descriptors [32], Gabor features [33], and Steerable Pyramidal Phase Features [34]. The last three comparators use feature vectors extracted by three Convolutional Neural Networks: VGG-Face [35], which has been trained for classifying faces, so the periocular region appears in the training data, and the very-deep Resnet101 [36] and Densenet201 [37] networks. Two example images from the two databases employed are shown in Fig. 3 (first column). The second column shows the two images after applying Contrast Limited Adaptive Histogram Equalization (CLAHE) [38], whereas the last two columns show the regions of interest (ROI) used by the different comparators. The comparators are evaluated both in terms of performance, template size and computation times. In a previous study [39], we presented preliminary results with the VSSIRIS database using a subset of the mentioned comparators [12,32,33], which are extended in the present paper with additional experiments using new comparators [34–37] and the mentioned Cross-Eyed database. Third, we describe our multialgorithm fusion architecture for periocular recognition using images from different sensors (Fig. 4). The input to a biometric comparator is usually a pair of biometric samples, and the output is, in general, a similarity score s. A larger score favours the hypothesis that the two samples come from the same subject (target or client hypothesis), whereas a smaller score supports the opposite (non-target or impostor hypothesis). However, if we consider a single isolated score from a biometric comparator (say a similarity score of $s = 1$), it is in general not possible to determine which is the hypothesis the score supports the most, unless we know the distributions of target or non-target scores. Moreover, since the scores output by the various comparators are heterogeneous, score normalization is needed to transform these scores into a common domain prior to the fusion process [20]. We solve these problems by linear logistic regression fusion [40,41], a trained classification approach in which scores of the individual comparators are combined to obtain a log-likelihood ratio. This is the logarithm of the ratio between the likelihood that input signals were originated by the same subject and the likelihood that input signals were not originated by the same subject. This form of output is comparator-independent in the sense that this log-likelihood-ratio output can theoretically be used to make optimal (Bayes) decisions. To convert scores from different comparators into a log-likelihood ratio, we evaluate two possibilities (Fig. 5). In the first one (top part), the mapping function uses as input the scores of all comparators, producing a single log-likelihood ratio as output. In the second one (bottom), several mapping functions are trained (one per comparator), so one log-likelihood ratio per comparator is obtained. Under independence assumptions (as in the case of comparators based on different feature extraction methods), the sum of log-likelihood ratios results in another log-likelihood ratio [42]. Therefore, in the second case, the outputs of the different mapping functions are just summed. The latter provides a simple fusion framework that allows obtaining a single log-likelihood ratio by simply summing the (mapped) score given by each available comparator. This would allow coping with missing modalities [43] since the output still would be a log-likelihood ratio regardless of the number of systems combined. This fusion approach has been previously

Table 1
Overview of existing works in periocular biometrics using images from different sensors. The works of each sub-section are in chronological order. The acronyms of this table are fully defined in the text.

Ref.	Features	Database	People/ Images	Best accuracy						
				Comparison	# Eyes	EER	GAR @ 1%FAR	GAR @ 0.1%FAR	GAR @ 0.01%FAR	Rank-1
Cross-sensor comparisons in the visible range (VIS–VIS)										
[45]	LBP, HOG, SIFT, ULBP, GIST	CSIP	50/2004	VIS–VIS	Single	15.5%	–	–	–	–
[46]	LD+STFT	MICHE I	50/n-a	VIS–VIS	Single	6.38–8.33%	–	–	–	–
[47]	GMM-UBM, SV-SDA, CNN	CSIP	50/2004	VIS–VIS	Single	–	–	–	–	83.6–93.3%
This work: 9 comparators		VSSIRIS	56/560	VIS–VIS	Single	0.3%	–	–	99.7%	–
Cross-sensor comparisons in the near-infrared range (NIR–NIR)										
[16]	OM	Own	300/9000	NIR–NIR	Single	20–28%	–	–	–	–
Cross-sensor comparisons across different spectra (cross-spectrum)										
[18]	LBP, NGC, JDSR	Own	704/1358	VIS–NIR	Single	23%	–	–	–	–
[48]	PHOG	IIITD-IMP	62/1240	VIS–NIR	Single/both	–	38.36/47.08%	–	–	–
				VIS-night	Single/both	–	63.81/71.93%	–	–	–
				NIR-night	Single/both	–	40.36/48.21%	–	–	–
[49]	Gabor+ WLD/LBP/HOG	Pre-Tinders	48/576	VIS-SWIR 1.5/50/106 m	Single	7.32/24.87/31.18%	–	–	–	68.75/33.33/31.94%
		Tinders	48/1255	VIS–NIR 1.5/50/106 m	Single	4.42/25.71/39.01%	–	–	–	70.31/38.54/10.76%
		PCSO	1000/3000	VIS-MWIR 1.5 m	Single	30.46%	–	–	–	5.58%
		Q-FIRE	82/431	VIS-LWIR 2 m	Single	39.06%	–	–	–	8.09%
[50]	MRF+ TPLBP/FPLB	IIITD IMP	62/1240	VIS–NIR	Single	–	–	15.93–18.35%	–	–
		PolyU	209/12540	VIS–NIR	Single	19.8–32.5%	–	45.4–73.2%	–	–
[51]	DOG+LBP/HOG	IIITD-IMP	62/1240	VIS–NIR	Single/both	43.85/45.29%	–	24.97/25.03%	–	–
		PolyU	209/12540	VIS–NIR	Single/both	18.79/13.87%	–	73.12/83.12%	–	–
		Cross-Eyed	120/3840	VIS–NIR	Single/both	15.11/10.36%	–	80.03/89.27%	–	–
[52]	HOG, GIST, LG, BSIF	Own	52/4160	8 bands	both	–	–	–	–	8.46–91.92%
[53]	CNN	IIITD-IMP	62/1240	VIS–NIR	Single	5.19%	88.13%	–	–	–
				VIS-night	Single	5.13%	88.19%	–	–	–
				NIR-night	Single	10.19%	81.55%	–	–	–
This work: 9 comparators		Cross-Eyed	120/3840	VIS–NIR	Single	0.2%	–	–	99.53%	–

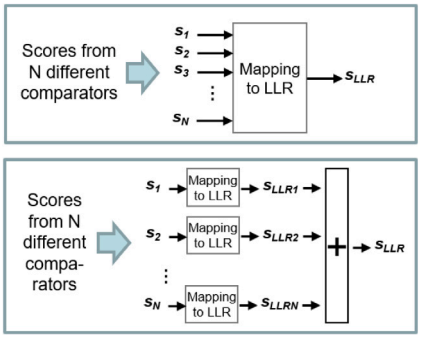

Fig. 5. Strategies to convert scores from multiple subsystems to a log-likelihood ratio (LLR). Top: one single mapping function is trained to convert multiple scores into a single LLR. Bottom: several mapping functions are trained to convert the score of each subsystem into a LLR. The sum of LLRs from different subsystems also results in a LLR. See the text for details.

applied successfully to cross-sensor comparison in the face and fingerprint modalities [23], achieving excellent results in other competition benchmarks as well [43]. Fourth, we compare this fusion approach with a set of simple and trained score fusion rules based on the arithmetic average of normalized scores [24], Support Vector Machines [25], and Random Forest [26]. These fusion approaches are very popular in the literature, having demonstrated to give good results in biometric authentication [20,44]. Fifth, in our experiments, conducted according to the 1st Cross-Spectral Iris/Periocular Competition (Cross-Eyed 2016) protocol [27], reductions of up to 29/47% in EER/FRR error rates (with respect to the best individual system) are obtained by fusion under NIR–VIS comparison, resulting in a cross-spectral EER of 0.2%, and a FRR @ FAR = 0.01% of just 0.47%. Regarding cross-sensor VIS–VIS smartphone recognition, the reductions in error rates achieved are 85/93% in EER/FRR, respectively, with corresponding cross-sensor error values of 0.3% (EER) and 0.3% (FRR).

2. Related works in periocular biometrics using images from different sensors

Interoperability between different sensors is an area of high research interest due to new scenarios arising from the widespread use of biometric technologies, coupled with the availability of multiple sensors and vendor solutions. A summary of existing works in the literature is given in Table 1. Most of them employ the Genuine Acceptance Rate (GAR) as metric, which is computed as 100-FRR(%). For this reason, in this subsection, we report GAR values. However, in the rest of the paper, we will follow the Cross-Eyed protocol and will report FRR values.

Cross-sensor comparison of images in the visible range (VIS–VIS) from smartphone sensors is carried out, for example, in [45–47], while the challenge of comparing images from different sensors in the near-infrared spectrum (NIR–NIR) has been addressed in [16]. In the work [46], the authors apply Laplacian decomposition (LD) of the image coupled with dynamic scale selection, followed by frequency decomposition via Short-Term Fourier Transform (STFT). In the experiments, they employ a subset of 50 periocular instances from the MICHE I dataset (Mobile Iris Challenge Evaluation I dataset) [54], captured with the front and rear cameras of two smartphones in indoor and outdoor illuminations. The cross-sensor EER obtained ranges from 6.38 to 8.33% for the different combinations of reference and probe cameras. The authors in [45] use a sensor-specific colour correction technique, which is estimated by using a colour chart in a dark acquisition scene that is further illuminated by a standard illuminant. The authors also carry out a score-level fusion of six iris and five periocular comparators, which is done by Neural Networks. The five periocular features include Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT) key-points, Uniform Local Binary Patterns (ULBP) [55], and the perceptual GIST descriptors [56]. They also presented a new database (CSIP: Cross-Sensor Iris and Periocular), with 2004 periocular images from 50 subjects captured with four different smartphones in ten different setups (based on several combinations involving the use of frontal/rear cameras and flash/no flash). The best reported periocular performance by fusion of the five available comparators is EER = 15.5%. The same database is also

employed in [47], where the authors apply three different methods to solve the cross-sensor task: Gaussian Mixture Models coupled with Universal Background Models (GMM-UBM), GMM Supervectors coupled with Stacked Denoising Autoencoders (SV-SDA), and deep transfer learning with Convolutional Neural Networks (CNN). They achieve a rank-1 recognition rate of 93.3% in the best possible case. The work [16], on the other hand, addresses the issue of cross-sensor recognition in the NIR spectrum. The authors employ a self-captured database with 9000 iris images from 600 eyes (300 people) using three different high-resolution sensors. Sensor interoperability is dealt with by weighted fusion of information from multiple directions of Ordinal Measures (OM), with a reported cross-sensor periocular EER between 20 and 28%.

Regarding recognition across different spectra (cross-spectral), the work [18] proposes to compare images of the periocular region cropped from VIS face images against NIR iris images. This is because face images are usually captured in the visible range, while iris images in commercial systems are usually acquired using near-infrared illumination. They employ three different comparators based on Local Binary Patterns (LBP), Normalized Gradient Correlation (NGC), and Joint Database Sparse Representation (JDSR). Using a self-captured database with 1358 images of the left eye from 704 subjects, they report a cross-spectral EER of 23% by score-level fusion of the three comparators.

In another line of work, surveillance at night or in harsh environments has prompted interest in new imaging modalities. For example, the authors in [48] presented the IIITD Multispectral Periocular database (IIITD-IMP), with a total of 1240 VIS, NIR and Night Vision images from 62 subjects (the latter captured with a video camera in Night Vision mode). To cope with cross-spectral periocular comparisons, they employ Neural Networks to learn the variabilities caused by each pair of spectra. The employed comparator is based on a Pyramid of Histograms of Oriented Gradients (PHOG) [57]. They report results for each eye separately and for the combination of both eyes, obtaining a cross-spectral GAR of 38%–64% at FAR = 1% (best of the two eyes), and a GAR of 47%–72% combining the two eyes. The use of pre-trained Convolutional Neural Networks (CNN) as a feature extraction method for NIR–VIS comparison was recently proposed in [53]. Here, the authors identify the layer of the ResNet101 network that provides the best performance on each spectrum. Then, they train a Neural Network that uses as input the feature vector of the best respective layers. Using the IIITD-IMP database, they report results considering the left and right eyes of a person as different users (effectively duplicating the number of classes). The obtained cross-spectral accuracy is EER = 5%–10% and GAR = 81%–88% at FAR = 1%, which outperforms any previous study with this database. The authors in [50] employ the IIITD-IMP database, and a newly presented database, the Hong Kong Polytechnic University Cross-Spectral Iris Images Database (PolyU), with 12 540 images from 209 subjects. To carry out NIR–VIS comparison, they use Markov Random Fields (MRF) combined with two different feature extraction methods, variants of Local Binary Patterns (LBP), namely FPLBP (Four-Patch LBP) and TPLBP (Three-Patch LBP). They report a cross-spectral periocular GAR at FAR = 0.1% of 16%–18% (IIITD-IMP) and 45%–73% (PolyU). These two databases, together with the Cross-Eyed database (with 3840 images in NIR and VIS spectra from 120 subjects) [27], are used in the work [51]. To normalize the differences in illumination between NIR and VIS images, they apply Difference of Gaussian (DoG) filtering. The comparators employed were based on Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG) features. They report results for each eye separately and for the combination of both eyes. The IIITD-IMP database gives the worst results, with a cross-spectral EER of 45% and a GAR at FAR = 0.1% of only 25% (two eyes combined). The reported accuracy with the other databases is better, ranging between 10%–14% (EER) and 83%–89% (GAR).

Latest advancements have resulted in devices with the ability to see through fog, rain, at night, and to operate at long ranges. In the work [49], the authors carry out experiments with several databases containing images with different wavelengths, namely VIS, NIR, SWIR (ShortWave Infrared), MWIR (MiddleWave Infrared), and LWIR (LongWave Infrared). The images are captured at several stand-off distances of 1.5 m, 2 m, 50 m, and 105 m. Feature extraction is done with a bank of Gabor filters, with the magnitude and phase responses further encoded with three descriptors: Weber Local Descriptor (WLD) [58], Local Binary Patterns (LBP), and Histogram of Oriented Gradients (HOG). Extensive experiments are done in this work comparing SWIR, NIR, MWIR and LWIR periocular probes to a gallery of VIS images. As expected, accuracy decreased as the standoff distance increases. Also, the comparison of MWIR or LWIR images to VIS images shows poor performance, attributable to the fact that MWIR and LWIR imagery measures the heat of a body, while visible imagery measures reflected light. Recently, the work [52] presented a new multispectral database captured in eight bands across the VIS and NIR spectra (530 to 1000 nm). A total of 4160 images from 52 subjects were acquired using a custom-built sensor that captures periocular images simultaneously in the eight bands. The comparators evaluated are based on Histogram of Oriented Gradients (HOG), perceptual descriptors (GIST), Log-Gabor filters (LG), and Binarized Statistical Image Features (BSIF). The cross-band accuracy varies greatly depending on the reference and probe bands, ranging from 8.46% to 91.92% rank-1 identification rate.

3. Periocular comparators

This section describes the biometric comparators used for periocular recognition. We employ nine different comparators, whose choice is motivated as follows. Three comparators are based on the most widely used features in periocular research, which are employed as a baseline in many studies [1]: Histogram of Oriented Gradients (HOG) [29], Local Binary Patterns (LBP) [30], and Scale-Invariant Feature Transform (SIFT) key-points [31]. Other three comparators, available in-house, have been self-developed by the authors and published previously with competitive results. These are based on Symmetry Descriptors (SAFE) [32], Gabor features (GABOR) [33], and Steerable Pyramidal Phase Features (NTNU) [34]. We also employ three comparators based on deep Convolutional Neural Networks: the VGG-Face network [35], which has been trained for classifying faces (so the periocular region appears in the training data), and the two very-deep Resnet101 [36] and Densenet201 [37] architectures.

3.1. Based on symmetry patterns (SAFE)

This comparator employs the Symmetry Assessment by Feature Expansion (SAFE) descriptor [32], which encodes the presence of various symmetric curve families around image key-points (Fig. 6, top). We use the eye centre as the anchor point for feature extraction. The algorithm starts by extracting the complex orientation map of the image via symmetry derivatives of Gaussians [59]. We employ $S = 6$ different scales in computing the orientation map, therefore capturing features at different scales, with standard deviation of each scale given by $\sigma_s = K^{s-1}\sigma_0$ (with $s = 1, 2..., S$; $K = 2^{1/3}$; $\sigma_0 = 1.6$). These parameters have been chosen according to [31]. For each scale, we then project $N_f = 3$ ring-shaped areas of different radii around the eye centre onto a space of $N_h = 9$ harmonic functions. We use the result of scalar products of complex harmonic filters (shown in Fig. 6) with the orientation image to quantify the amount of presence of different symmetric pattern families within each annular band. The resulting complex feature vector is given by an array of $S \times N_h \times N_f$ elements. The comparison score $M \in \mathbb{C}$ between a query q and a test SAFE array t is computed using the triangle inequality as $M = \frac{\langle q,t \rangle}{\langle |q|,|t| \rangle}$. The argument $\angle M$ represents the angle between the two arrays (expected to be zero when the symmetry patterns detected coincide for reference

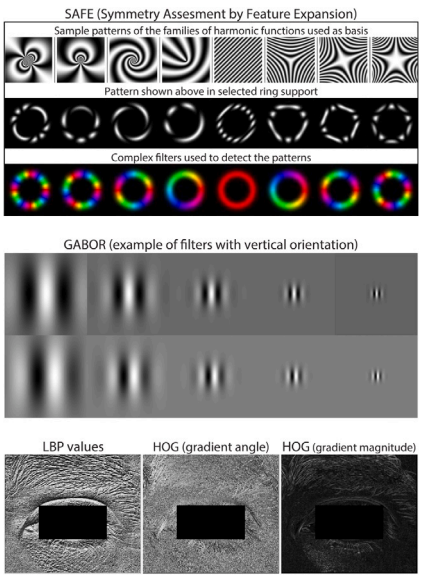

Fig. 6. Example of some feature extraction methods employed. **SAFE comparator.** Example of symmetric curve families and complex filters used to detect the patterns. Hue in colour images encode the direction, and saturation represents the complex magnitude. **GABOR comparator.** Gabor filters with vertical orientation (top: real part, bottom: imaginary part). Depicted filters are of size 88 × 88, with wavelengths spanning logarithmically the range from 44 (first column) to 6 pixels (last column). **LBP and HOG comparators.** Example of LBP and HOG features of the input image shown in Fig. 3 (top row). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and test feature vectors, and 180° when they are orthogonal), and the confidence is given by $|M| \in [0,1]$. To include confidence into the angle difference, we use $MS = |M| \cos \angle M$, with the resulting score $MS \in [-1,1]$.

The annular band of the first ring is set in proportion to the distance between eye corners (Cross-Eyed database) or to the radius of the sclera circle (VSSIRIS database), while the band of the last ring ends at the boundary of the image. This difference in setting the smallest ring is due to the ground-truth information available for each database, as explained later. However, in setting the origin of the smallest band, we have tried to ensure that the different annular rings capture approximately the same relative spatial region in both databases. The ROI of the SAFE comparator for each database is shown in Fig. 3 (third column). Using the eye corners or the sclera boundary as reference for the first annular band alleviates the effect of dilation that affects the pupil, which is more pronounced with visible illumination. Since the eye corners or the sclera are not affected by such dilation or by partial occlusion due to eyelids, they provide a more stable Ref. [60].

3.2. Based on Gabor features (GABOR)

This comparator is described in [33], which is based on the face recognition comparator presented in [61]. The periocular image is decomposed into non-overlapped square regions (Fig. 3, fourth column), and the local power spectrum is then sampled at the centre of each block by a set of Gabor filters organized in 5 frequency and 6 orientation channels. An example of Gabor filters is shown in Fig. 6. This sparseness of the sampling grid allows direct Gabor filtering in the image domain without needing the Fourier transform, with significant computational savings and feasibility in real-time. Gabor responses from all grid points are grouped into a single complex vector, and the comparison between two images is made using the magnitude of complex values via the χ^2 distance. Prior to the comparison with magnitude vectors, they are normalized to a probability distribution (PDF). The χ^2 distance between a query q and a test vector t is computed as $\chi^2_{qt} = \sum_{n=1}^{N} \frac{(p_q[n]-p_t[n])^2}{p_q[n]+p_t[n]}$, where p are entries in the PDF, n is the bin index, and N is the number of bins in the PDF (dimensionality). The χ^2 distance, due to the denominator, gives more weight to low probability regions of the PDF. For this reason, it has been observed to produce better results than other distances when using normalized histograms [62].

3.3. Based on Steerable Pyramidal Phase Features (NTNU)

Image features from multi-scale pyramids have proven to extract discriminative features in many earlier works concerned with texture synthesis, texture retrieval, image fusion, and texture classification, among others [63–70]. Inspired by this applicability, we employ steerable pyramidal features for periocular image classification using images from different sensors. Further, observing the nature of textures that are different across spectra (NIR versus VIS), we propose to employ the quantized phase information from the multi-scale pyramid of the image, as explained next.

A steerable pyramid is a translation and rotation invariant transform in a multi-scale, multi-orientation and self-inverting image decomposition into a number of sub-bands [71–73]. The pyramidal decomposition is performed using directional derivative operators of a specific order. The key motivation in using steerable pyramids is to obtain both linear and shift-invariant features in a single operation. Further, they not only provide multi-scale decomposition but also provide the advantages of orthonormal wavelet transforms that are both localized in space and spatial-frequency with aliasing effects [71]. The basis functions of a steerable pyramid are K-order directional derivative operators. The steerable pyramids come in different scales and $K+1$ orientations.

For a given input image, the features of steerable pyramid coefficients can be represented using $S_{(m,\theta)}$, where m represents the scale and θ represents the orientation. In this work, we generate a steerable pyramid with 3 scales ($m \in \{1,2,3\}$) and angular coefficients in the range $\theta_1 = 0$ to $\theta_{K+1} = 360$, resulting in a pyramid that covers all directions. The set of sub-band images corresponding to one scale can be therefore represented as $S_m = \{S_{(m,\theta_1)}, S_{(m,\theta_2)}, \ldots, S_{(m,\theta_{K+1})}\}$. We further note that the textural information represented is different in the NIR and VIS domains. In order to obtain domain invariant features, we propose to extract the local phase features [74] from each sub-band image $S_{(m,\theta)}$ in a local region ω in the neighbourhood of n pixels given by $F_{(m,\theta)}(u,x) = S_{(m,\theta)}(x,y)\omega_R(y-x)\exp\{-j2\pi U^T y\}$, where x, y represent the pixel location. The local phase response obtained through Fourier coefficients are computed for the frequency points u_1, u_2, u_3 and u_4, which relate to four points $[a,0]^T, [0,a]^T, [a,a]^T, [a,-a]^T$ such that the phase response $H(u_i) > 0$ [74]. The phase information presented in the form of Fourier coefficients is then separated into real and imaginary parts of each component, as given by $[Re\{F\}, Im\{F\}]$, to form a vector R with eight elements. Next, the elements R_i of R are binarized to Q_i by assigning a value of 1 to components with a response greater than 1, and 0 otherwise. The phase information is finally encoded to a compact pixel representation P in the $0-255$ range by using a simple binary to decimal conversion strategy given by $P_{(m,\theta)} = \sum_{j=1}^{8} Q_j \times (2^{(j-1)})$.

This procedure is followed with the different scales and orientations of the selected space. All the phase responses $P_{(m,\theta)}$ of the input image are concatenated into a single vector. Comparison between feature representations of two images is made using the χ^2 distance.

3.4. Based on SIFT key-points (SIFT)

This comparator is based on the SIFT operator [31]. SIFT key-points (with dimension 128 per key-point) are extracted in the annular ROI shown in Fig. 3, third column. The use of an annular ROI like SAFE is inherited from our previous contribution [39], but to compare with other systems that employ the entire input image (Fig. 3, fourth column), we report experiments with the latter as well. The recognition metric between two images is the number of paired key-points, normalized

by the minimum number of detected key-points in the two images being compared. We use a free C++ implementation of the SIFT algorithm,[1] with the adaptations described in [75]. Particularly, it includes a post-processing step to remove spurious pairings using geometric constraints, so pairs whose orientation and length differ substantially from the predominant orientation and length are removed.

3.5. Based on Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG)

Together with SIFT key-points, LBP [30] and HOG [29] have been the most widely used descriptors in periocular research [1]. An example of LBP and HOG features is shown in Fig. 6, bottom. The periocular image is decomposed into non-overlapped regions, as with the Gabor comparator (Fig. 3, fourth column). Then, HOG and LBP features are extracted from each block. Both HOG and LBP are quantized into 8 different values to construct an 8 bins histogram per block. Histograms from each block are then normalized to account for local illumination and contrast variations and finally concatenated to build a single descriptor of the whole periocular region. Image comparison with HOG and LBP can be made by simple distance measures. Euclidean distance is usually used for this purpose [12], but here we employ the χ^2 distance for the same reasons as with the Gabor comparator.

3.6. Based on deep convolutional Neural Networks (VGG-face, Resnet101, Densenet201)

Inspired by the works [53,76,77] in iris and periocular biometrics, we leverage the power of existing architectures pre-trained with millions of images to classify hundreds of thousands of object categories.[2] They have proven to be successful in very large recognition tasks apart from the detection and classification tasks for which they were designed [78].

Here, we employ the VGG-Face [35] and the very deep Resnet101 [36] and Densenet201 [37] architectures. VGG-Face is based on the VGG-Very-Deep-16 CNN sequential architecture, implemented using ~1 million images from the Labelled Faces in the Wild [79] and YouTube Faces [80] datasets. Since VGG-Face is trained for classifying faces, we believe that it can provide effective recognition with the periocular region as well, given that this region appears in the training images. Introduced later, the ResNet networks [36] presented the concept of residual connections to ease the training of CNNs. By reducing the number of training parameters, they can be substantially deeper. The key idea of residual connections is to make available the input of a lower layer to a higher layer, bypassing intermediate ones. There are different variants of ResNet networks, depending on its depth. In this work, we employ ResNet101, having a depth of 347 layers (including 101 convolutional layers). In DenseNet networks [37], the residual concept is taken even further since the feature maps of all preceding layers of a Dense block are used as inputs of a given layer, and its own feature maps are used as inputs into all subsequent layers. This encourages feature reuse throughout the network. Similarly to ResNet, there are different variants of DenseNet (defined by its depth). In this paper, we employ Densenet201, having a depth of 709 layers (including 201 convolutional layers).

In using these networks, periocular images are fed into the feature extraction pipeline of each pre-trained CNN [76,77]. However, instead of using the vector from the last layer, we employ as feature descriptor the vector from the intermediate layer identified as the one providing the best performance. These will be found in the respective experimental sections. This approach allows the use of powerful architectures pre-trained with a large number of images in a related domain, eliminating the need of designing or re-training a new network for a specific task, which may be infeasible in case of lack of large-scale databases in the target domain (as in the case of periocular recognition with images from different sensors). The extracted CNN vectors can be simply compared with distance measures. In our case, we employ the χ^2 distance, which has proven to provide better results than other measures such as the cosine or Euclidean distances [77].

[1] http://vision.ucla.edu/~vedaldi/code/sift/assets/sift/index.html.

[2] ImageNet. http://www.image-net.org.

4. Score fusion methods

A biometric verification comparator can be defined as a pattern recognition machine that, by comparing two (or more) samples of input signals, is designed to recognize two different classes. The two hypotheses or classes defined for each comparison are *target* hypothesis (θ_t: the compared biometric data comes from the same individual) and *non-target* hypothesis (θ_{nt}: the compared data comes from different individuals). As a result of the comparison, the biometric system outputs a real number s known as *score*. The higher the score, the more it supports the target hypothesis, and vice-versa. The acceptance or rejection of an individual is based on a decision threshold τ, and this threshold depends on the priors and decision costs involved in the decision-making process. However, if we do not know the distributions of target or non-target scores from such comparator or any threshold, we will not be able to classify the associated biometric samples in general.

Integration at the score level is the most common approach used in multibiometric systems due to the ease in accessing and combining the scores $\mathbf{s} = (s_1, \ldots, s_i, \ldots, s_N)$ generated by N different comparators [20]. Unfortunately, each biometric comparator outputs scores that are in a range that is specific to the comparator, so score normalization is needed to transform these scores into a common domain prior to the fusion [24], e.g. $s_i \in [0, 1]$ or $s_i \in [-1, 1]$, $\forall i \in \{1, \ldots, N\}$. But even if two comparators output scores in the same range, the same output value (say $s_i = s_j = 0.5$ for $i \neq j$) might not favour the target or non-target hypotheses with the same strength. The same can be said about the fusion of such scores. From this viewpoint, outputs are dependent on the comparator, and thus, the acceptance/rejection decision also depends on the comparator.

These problems can be addressed with the concept of *calibrated* scores. During calibration, the scores $\mathbf{s} = (s_1, \ldots, s_i, \ldots, s_N)$ are mapped to a log-likelihood-ratio (LLR) as $s^{cal} \approx \log\left(\frac{p(\mathbf{s}|\theta_t)}{p(\mathbf{s}|\theta_{nt})}\right)$, where s^{cal} represents the calibrated score. Then, a decision can be taken using the Bayes decision rule [42]:

$$\text{For a given } \mathbf{s} \begin{cases} \text{decide } \theta_t : (p(\mathbf{s}|\theta_t)/p(\mathbf{s}|\theta_{nt})) > \tau_B \\ \text{decide } \theta_{nt} : (p(\mathbf{s}|\theta_t)/p(\mathbf{s}|\theta_{nt})) < \tau_B \end{cases} \quad (1)$$

The parameter τ_B is known as the *Bayes threshold*, and its value depends on the prior probabilities of the hypotheses $p(\theta_t)$ and $p(\theta_{nt})$ and on the decision costs. This form of output is *comparator-independent* since this log-likelihood-ratio output can theoretically be used to make optimal (Bayes) decisions for any given target prior and any costs associated with making erroneous decisions [42]. Therefore, the calibration process gives *meaning* to s^{cal}. In a Bayesian context, a calibrated score s^{cal} can be interpreted as a degree of support to any of the hypotheses. If $s^{cal} > 0$, then the support to θ_t is also higher, and vice-versa. Also, the meaning of a log-likelihood ratio is the same across different biometric comparators, allowing to compare them in the same probabilistic range. This calibration transformation then solves the two previously commented problems. First, it maps scores from biometric comparators to a common domain. Second, it allows the interpretation of biometric scores as a degree of support.

A number of strategies can be used to train a calibration transformation [81]. Among them, logistic regression has been successfully used for biometric applications [23,40,41,82,83]. With this method, the scores of multiple comparators are fused together, primarily to improve the discriminating ability, in such a way as to encourage good

calibration of the output scores. Given N biometric comparators which output the scores $\mathbf{s}_j = (s_{1j}, s_{2j}, \dots, s_{Nj})$ for an input trial j, a linear fusion of these scores is:

$$f_j = a_0 + a_1 \cdot s_{1j} + a_2 \cdot s_{2j} + \cdots + a_N \cdot s_{Nj} \tag{2}$$

When the weights $\{a_0, \dots, a_N\}$ are trained via logistic regression, the fused score f_j is a well-calibrated log-likelihood-ratio [41,81]. Let $[s_{ij}]$ be an $N \times N_T$ matrix of training scores built from N biometric comparators and N_T target trials, and let $[r_{ij}]$ be an $N \times N_{NT}$ matrix of training scores built from the same N biometric comparators with N_{NT} non-target trials. We use a logistic regression objective [40,41] that is normalized with respect to the proportion of target and non-target trials (N_T and N_{NT}, respectively), and weighted with respect to a given prior probability $P = P(\text{target})$. The objective is stated in terms of a *cost* C, which must be *minimized*:

$$C = \frac{P}{N_T} \sum_{j=1}^{N_T} \log\left(1 + e^{-f_j - \text{logit}\, P}\right) + \frac{1-P}{N_{NT}} \sum_{j=1}^{N_{NT}} \log\left(1 + e^{-g_j - \text{logit}\, P}\right) \tag{3}$$

where the fused target and non-target scores are respectively

$$f_j = a_0 + \sum_{i=1}^{N} a_i s_{ij}$$
$$g_j = a_0 + \sum_{i=1}^{N} a_i r_{ij} \tag{4}$$

and where $\text{logit}\, P = \log\left(\frac{P}{1-P}\right)$.

It can be demonstrated that minimizing the objective C with respect to $\{a_0, \dots, a_N\}$ results in a good calibration of the fused scores [41,81]. In practice, changing the value of P has a small effect. The default of 0.5 is a good choice for a general application and it will be used in this work. The optimization objective C is convex and therefore has a unique global minimum.

Another advantage of this method is that when fusing scores from different comparators, the most reliable comparator will implicitly be given a dominant role in the fusion (via the trained weights $\{a_0, \dots, a_N\}$). In other standard fusion methods, such as the average of scores [24], all comparators are given the same weight in the fusion, regardless of its individual accuracy. It is also straightforward to show that if M calibrated scores $\{s_1^{cal}, s_2^{cal}, \dots, s_M^{cal}\}$ come from statistically independent sources (such as multiple biometric comparators), its sum $s_1^{cal} + s_2^{cal} + \cdots + s_M^{cal}$ also yields a log-likelihood ratio [42]. The latter allows to calibrate the scores s_i of each available biometric comparator separately (by using $N = 1$ in Eq. (2)), and simply sum the calibrated scores s_i^{cal} of each comparator in order to obtain a new calibrated score, as shown in Fig. 5. In this paper, the two possibilities are evaluated, i.e. calibrating the scores of all comparators together vs.calibrating them separately and then summing them up. In order to perform logistic regression calibration, the freely available Bosaris toolkit for Matlab has been used.[3] For further details of this fusion method, the reader is referred to [23] and the references therein.

The probabilistic fusion method described above is compared in the present work with three strategies. Since each biometric comparator usually outputs scores that are in a range that is specific to the system, the scores of each comparator are normalized prior to the fusion using z-score normalization [24]. The three strategies are:

- **Average**. With this simple rule, the scores of the different comparators are simply averaged. Motivated by their simplicity, simple fusion rules have been used in biometric authentication with very good results [84,85]. They have the advantage of not needing training, sometimes surpassing other complex fusion approaches [86].
- **SVM**. Here, a Support Vector Machine (SVM) is trained to provide a binary classification given a set of scores from different biometric comparators [87]. The SVM algorithm searches for an optimal hyperplane that separates the data into two classes. SVM is a popular approach employed in multibiometrics [25], which has shown to outperform other trained approaches [20]. In this work, we evaluate Linear, RBF, and Polynomial (order 3) kernels. Instead of using the binary predicted class label, we use the signed distance to the decision boundary as the output score of the fusion. This allows the presentation of DET curves and associated EER and FRR measures.
- **Random Forest**. Another method employed for the fusion of scores from multiple biometric comparators is the Random Forest (RF) algorithm [26]. An extension of the standard classification tree algorithm, the RF algorithm is an ensemble method where the results of many decision trees are combined [88]. This helps to reduce overfitting and to improve generalization capabilities. The trees in the ensemble are grown by using bootstrap samples of the data. In this work, we evaluate ensembles with 25, 150, and 600 decision trees. Instead of using the binary predicted class label, we use the weighted average of the class posterior probabilities over the trees that support the predicted class, so we can present DET curves and associated measures.

[3] https://sites.google.com/site/bosaristoolkit/.

5. Cross-spectral (NIR-VIS) periocular recognition

5.1. Database and protocol

In the cross-spectral recognition experiments of this section, we employ the Reading Cross-Spectral Iris/Periocular Dataset used as the benchmark dataset for the 1st Cross-Spectral Iris/Periocular Competition (Cross-Eyed 2016) [27]. The dataset contains both visible (VIS) and near-infrared (NIR) images captured with a custom dual spectrum imaging sensor which acquires images in both spectra synchronously. Periocular images are of size 800 × 900 (height × width) from 120 subjects, with 8 images of both eyes captured in both spectra, totalling 3840 images. Images are captured at a distance of 1.5 m, in an uncontrolled indoor environment, containing large variations in ethnicity, eye colour, and illumination reflections. Some examples are shown in Fig. 7 (top). To avoid usage of iris information by periocular methods during the Cross-Eyed competition, periocular images were distributed with a mask on the eye region, as discussed in [12]. A new edition of the competition was held in 2017. The 120 subjects of the Cross-Eyed 2016 database were provided as the training set, and an additional set of 55 subjects were sequestered as the test set in the 2017 edition, but the latter was never released [89].

Prior to the competition, a *training* set of images from 30 subjects was distributed. The *test* set consisted of images from 80 subjects, sequestered by the organizers and distributed after the competition. Images from 10 additional subjects were also released after the competition that were not present in the test set. Here, we will employ the same 30 subjects of the training set to tune our algorithms and the remaining 90 subjects for testing purposes. All images have an annotation mask of the eye region. The mass centre of the mask is set as the reference point (centre) of the eye. Images are then rotated w.r.t. the axis that crosses the two mask corners and resized via bicubic interpolation to have the same corner-to-corner distance (set to 318 pixels, the average value of the training set). Then, images are aligned by extracting a region of 613 × 701 around the eye. This size is set empirically to ensure that all available images have sufficient margin to the four sides of the eye centre. Eyes in the Cross-Eyed database are slightly displaced in the vertical direction, so the eye is not centred in the aligned images but with a vertical offset of 56 pixels (see Fig. 3, top). Images are further processed by Contrast Limited Adaptive Histogram Equalization (CLAHE) [38], which is the preprocessing choice with ocular images [90], and then sent to feature extraction.

(a) Cross-Eyed database (top row: visible images, bottom: near-infrared)

(b) VSSIRIS database (top row: Apple iPhone 5S, bottom: Nokia Lumia 1020; images taken from [28])

Fig. 7. Sample periocular images. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

We carry out verification experiments, with each eye considered a different user. We compare images both from the same device (*same-sensor*) and from different devices (*cross-spectral*). Genuine trials are obtained by comparing each image of an eye to the remaining images of the same eye. In *same-sensor* comparisons, to avoid symmetric comparisons, the first image of an eye is compared to the second to eight images; the second image is compared to the third to eight images, and so on, leading to $(7 + 6 + ... + 1)$ genuine scores per eye. This procedure is repeated for the two eyes of all subjects. This results in 30 subjects $\times$ 2 eyes $\times$ $(7 + 6 + ... + 1)$ and 90 $\times$ 2 $\times$ $(7 + 6 + ... + 1)$ genuine scores with the training and test set, respectively. In *cross-spectral* comparisons, the eight images of an eye in one spectrum are compared against the eight images in the other spectrum, leading to 8 $\times$ 8 genuine scores per eye. This results in 30 subjects $\times$ 2 eyes $\times$ 8 $\times$ 8 and 90 $\times$ 2 $\times$ 8 $\times$ 8 genuine scores with the training and test set, respectively. Impostor trials are done by comparing the 1st image of an eye to the 2nd image of the remaining eyes. In *same-sensor* comparisons, given a subject of the test set, his/her 1st image of both eyes is compared against the 2nd image of both eyes from the remaining 89 subjects. This results in 89 $\times$ 4 test impostor scores per subject and 90 $\times$ 89 $\times$ 4 impostor scores in total. In *cross-spectral* comparisons, the number of impostor scores is doubled by comparing the 1st image in VIS against the 2nd image in NIR, and the 1st image in NIR against the 2nd image in VIS. This results in 90 $\times$ 89 $\times$ 4 $\times$ 2 test impostor scores in total. To increase the number of available training scores, we carry out an additional comparison to the 3rd image of the remaining eyes only with the training set, effectively duplicating the number of impostor scores per subject. Since the training set contains 30 subjects, this results in 29 $\times$ 4 $\times$ 2 (*same-sensor*) and 29 $\times$ 4 $\times$ 2 $\times$ 2 (*cross-spectral*) training impostor scores per subject. By multiplying these amounts by 30, we obtain the total amount of impostor scores with the training set. The experimental protocol is summarized in Table 2.

The periocular comparators employed have some parameters which are set as follows. It should be highlighted that these parameters are computed in proportion to the size of the image, without any other training. If the image size changed, they would adapt dynamically so that the comparators would always be capturing their features in the same relative areas of the image. The only input needed is the position of the eye corners, which were also used to normalize and crop the image to a constant size, as described above. Regarding the SAFE comparator, the annular band of the first circular ring starts at a radius of R = 79 pixels (determined empirically as 1/4 of the eye corner-to-corner distance), and the band of the last ring ends at the bottom boundary of the image. This results in a ROI of 501 $\times$ 501 pixels around the eye centre (as shown in Fig. 3, third column). The grid employed with GABOR, LBP and HOG comparators has 7 $\times$ 8 = 56 non-overlapping blocks. Based on the size of the input image, each block has 88 $\times$ 88 pixels. The 8 central blocks are not considered since they are equal for all users due to the eye region mask, so features are extracted only from 48 blocks. The GABOR comparator employs filter wavelengths spanning from 44 to 6 pixels, which are set proportional to the block size as 88/2 = 44 to 88/16≈6. The VGG-Face, Resnet101, and Densenet201 comparators employ an input image size of 224 $\times$ 224, so images are resized to match these dimensions. Regarding the SIFT comparator, our baseline configuration entails the use of the same annular ROI than the SAFE comparator [39] for key-point extraction. However, for comparison purposes with the other systems, we also evaluate the use of the entire input image (except the 8 central blocks). This is done both at the original size of the image (613 $\times$ 701) and at the input size of the CNNs (224 $\times$ 224). Table 3 (second column) indicates the size of the feature vector for a given periocular image with the different comparators employed. Obviously, the SIFT descriptor is dependent on the size of the ROI and the image. With the full ROI, the average number of key-points per image is 2543 (of which a vector of 128 elements is computed, resulting in 2543 $\times$ 128 = 325 504 real values per image). The annular ROI produces a slightly smaller amount (1900 key-points, or 243 200 values) and if the image is reduced to 224 $\times$ 224, the amount is substantially less (only 92 key-points, or 11 776 values). Experiments have been done in a Dell Latitude E7240 laptop with an i7-4600 (2.1 GHz) processor, 16 Gb DDR3 RAM, and a built-in Intel HD Graphics 4400 card. The OS is Microsoft Windows 8.1 Professional, and the comparators are implemented in Matlab x64, with the exception of SIFT that is implemented in C++ and invoked from Matlab via MEX files. The VGG-Face model is from Caffee, which has been imported to Matlab with the `importCaffeNetwork` function. The Resnet101 and Densenet201 models are from the pre-trained models available in Matlab r2019a. In line with the Cross-Eyed competition, we also provide the extraction and comparison time of each method (Table 4, second and third columns). Here, it can be also appreciated the variation of the SIFT versions depending on the image or ROI size.

5.2. Results: Finding the optimum layer of the convolutional Neural Networks

Normalized periocular images are fed into the feature extraction of each pre-trained CNN. We investigate the representation capability of each layer by reporting the corresponding cross-spectral accuracy using features from each layer. The recognition accuracy of each network (EER and FRR @ FAR = 0.01%) is given in Fig. 8. It is worth noting that the best performance is obtained in some intermediate layer for all CNNs, in line with previous studies using ocular modalities [76,77]. In selecting the best layer, we prioritize the FRR @ FAR = 0.01%, since this was the metric employed to rank submissions to the Cross-Eyed competition, although we seek a balance with the EER as well. We have also searched for layers that give optimum performance both with the Cross-Eyed and the VSSIRIS databases simultaneously if possible (results with the latter are given in Fig. 12).

A good performance with VGG-Face is obtained with layer 25, which is a max pooling layer with 14×14×512 = 100352 elements. Layer 27 also provides good performance. This is a ReLu layer of the same size as layer 25, but since it has many elements set to 0 due to the ReLu operation, we prefer to choose layer 25. VGG-Face is a serial network, with layers arranged one after the other. On the other hand, ResNet101 and Densenet201 are acyclic networks, in which layers have inputs from multiple layers and outputs to multiple layers. This more intricate architecture may thus explain the oscillations observed between layers.

Table 2
Cross-Eyed database: Experimental protocol. E = Eyes, L = Left eye, R = Right eye, S = Sensors.

Cross-Eyed database			
Comparison type		Training (30 subjects)	Test (90 subjects)
Same-Sensor	Genuine	$30 \times 2E \times (7 + 6 + \cdots + 1) = 1680$	$90 \times 2E \times (7 + 6 + \cdots + 1) = 5040$
	Impostor	$30 \times 29 \times (4L + 4R) = 6960$	$90 \times 89 \times (2L + 2R) = 32{,}040$
Cross-Spectral	Genuine	$30 \times 2E \times 8L \times 8R = 3840$	$90 \times 2E \times 8L \times 8R = 11{,}520$
	Impostor	$30 \times 29 \times (4L + 4R) \times 2S = 13{,}920$	$90 \times 89 \times (2L + 2R) \times 2S = 64{,}080$

Table 3
Size of the feature vector per comparator and per database. AR = annular ROI. FR = Full ROI. 'Original' refers to the original size of the input image (Cross-Eyed: 613 × 701, VSSIRIS: 871 × 871).

Comparator	Cross-Eyed	VSSIRIS	Data
SAFE	$6 \times 3 \times 9 = 162$	$6 \times 3 \times 9 = 162$	Complex
GABOR	$48 \times 30 = 1440$	$56 \times 30 = 1680$	Real
SIFT (AR original)	circa 243200	circa 384000	Real
SIFT (FR original)	circa 325504	circa 489472	Real
SIFT (FR 224 × 224)	circa 11776	circa 16512	Real
LBP, HOG	$48 \times 8 = 384$	$56 \times 8 = 448$	Real
NTNU	9472	9472	Integer
VGG-Face	100 352	100 352	Real
Resnet101	50 176	100 352	Real
Densenet201	6272	43 904	Real

Table 4
Feature computation times for each database. AR = annular ROI. FR = Full ROI. 'Original' refers to the original size of the input image (Cross-Eyed: 613 × 701, VSSIRIS: 871 × 871).

	Cross-Eyed database		VSSIRIS database	
	Extraction time	Comparison time	Extraction time	Comparison time
SAFE	2.98 s	0.2 ms	11.86 s	<0.1 ms
GABOR	0.49 s	0.3 ms	0.53 s	0.3 ms
SIFT (AR original)	0.94 s	0.58 s	1.5 s	1.1 s
SIFT (FR original)	0.94 s	0.94 s	1.5 s	1.7 s
SIFT (FR 224 × 224)	0.05 s	1.6 ms	0.07 s	3 ms
LBP	0.16 s	<0.1 ms	0.17 s	<0.1 ms
HOG	0.01 s	<0.1 ms	0.13 s	<0.1 ms
NTNU	0.6 s	0.7 ms	0.56 s	0.7 ms
VGG-Face	0.51 s	1.65 ms	0.52 s	1.43 ms
Resnet101	0.27 s	0.35 ms	0.48 s	0.65 ms
Densenet201	0.25 s	<0.1 ms	0.39 s	0.42 ms

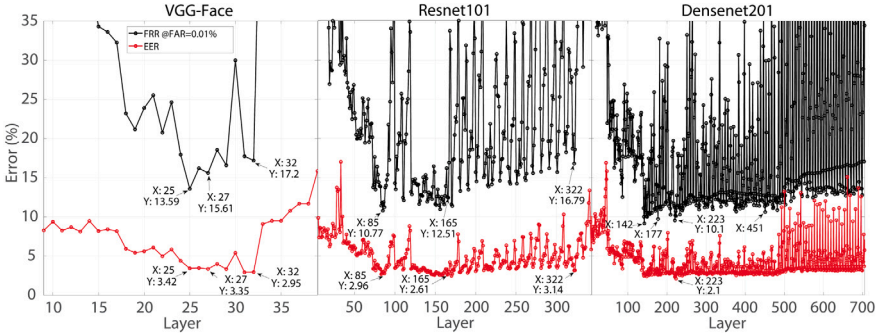

Fig. 8. Cross-Eyed database: Cross-spectral accuracy (VIS–NIR) of different CNN layers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

With ResNet101, a good performance is obtained with layer 165. This is a convolutional layer with $14 \times 14 \times 256 = 50176$ elements, and it will be the layer employed with Cross-Eyed. With VSSIRIS, better performance is obtained with layer 323, which is not the case with Cross-Eyed. This is a ReLu layer with $7 \times 7 \times 2018 = 100352$ elements. We choose this layer with VSSIRIS instead since it provides better EER than other layers as well. Regarding DenseNet201, good performance with Cross-Eyed (which minimizes both the EER and FRR) is obtained with layer 223. This is a convolutional layer with only $14 \times 14 \times 32 = 6272$ elements. Other layers (e.g. 142 or 177) also give a good FRR, but the

Table 5
Cross-Eyed database, test set: Verification results of the individual comparators. The relative variation of cross-spectral performance with respect to the best same-sensor performance is given in brackets (for the SIFT rows with VIS = 0%, the result is calculated w.r.t. the NIR performance to avoid division by zero). AR = annular ROI. FR = Full ROI.

Comparator	Equal Error Rate (EER)			FRR @ FAR = 0.01%		
	Same sensor		Cross-spectral	Same sensor		Cross-spectral
	NIR	VIS		NIR	VIS	
SAFE	5.85%	5.67%	9.47% (+67%)	22.4%	24.23%	50.38% (+124.9%)
GABOR	5.48%	5.34%	7.94% (+48.7%)	26.25%	23.68%	43.3% (+82.9%)
SIFT (AR 613 × 701)	0.02%	0%	0.28% (>1300%)	0.02%	0%	0.88% (>4300%)
SIFT (FR 613 × 701)	0.02%	0%	0.27% (>1250%)	0.02%	0%	0.9% (>4400%)
SIFT (FR 224 × 224)	1.11%	0.86%	3.36% (+290%)	5.83%	2.98%	29.7% (+897%)
LBP	3.03%	3.27%	5.84% (+92.7%)	10.97%	12.86%	63.79% (+481.5%)
HOG	3.84%	4.19%	5.06% (+31.8%)	11.76%	14.93%	34.36% (+192.2%)
NTNU	2.83%	2.45%	4.22% (+72.2%)	3.93%	3.57%	13.8% (+286.6%)
VGG-Face	2.36%	2.53%	3.42% (+44.9%)	8.48%	8.68%	13.59% (+60.3%)
Resnet101	1.52%	1.6%	2.61% (+71.7%)	5.51%	5.01%	12.51% (+149.7%)
Densenet201	1.37%	1.54%	2.09% (+52.6%)	5.69%	5.18%	10.09% (+94.8%)

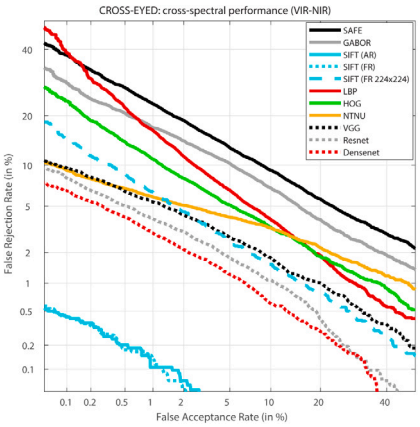

Fig. 9. Cross-Eyed database, test set: Verification results of the individual comparators. Best seen in colour. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

EER is not as good as with layer 223. With VSSIRIS, better performance is given by layer 480 instead, which is an average pooling layer with $7 \times 7 \times 896 = 43904$ elements.

5.3. Results: Individual comparators

We now report the performance of all periocular comparators in Table 5. Besides the EER, we also report the FRR at FAR = 0.01%. The latter was the metric used to rank submissions to the Cross-Eyed competition. We report two types of results: (*i*) *same-sensor* comparisons; and (*ii*) *cross-spectral* comparisons. In Fig. 9 we also give the DET curves of the cross-spectral experiments.

From Table 5, it can be seen that given a comparator, the NIR and VIS performances (same-sensor) are relatively equal. For example, the EER of SAFE is 5.85% (NIR) and 5.67% (VIS), so if the two images are in the same spectrum, there is no significant advantage in operating in NIR or VIS. This happens with all comparators, both in the EER and the FRR (with just a few exceptions), which is very interesting because they are based on different image features. In previous studies, the periocular modality usually performed better with VIS data [91–93], so it is generally accepted this modality is most suited to VIS imagery [1]. On the contrary, some other works show opposite results [48]. However, in the mentioned studies, the images employed are of smaller size, ranging from 100 × 160 to 640 × 480, while the images employed in this paper are of 613 × 701 pixels. Also, they evaluate three different periocular comparators at most. In the present paper, the use of bigger images may be the reason for a comparable performance between NIR and VIS images.

Regarding cross-spectral experiments, we observe a significant worsening in performance w.r.t. same-sensor comparisons, although not all comparators are affected in the same way. HOG, NTNU and especially LBP are substantially affected in high security mode (i.e. low FAR), as can be appreciated in the right part of Table 5. The relative FRR increase @ FAR = 0.01% for these comparators is in the range of 200% to nearly 500%. But the comparator that is most affected is SIFT. Even if its cross-spectral performance is the best among all comparators, it is about one or two orders of magnitude worse than its same-sensor performance (meaning a thousand per cent worse or more). This is despite the use of a descriptor with a bigger size (see Table 3). SIFT extracts features from a discrete set of local key-points, but it might be that the position of detected key-points is not the same in each spectrum. With the other comparators, on the other hand, the image is divided into annular or square regions (Fig. 3), and features are extracted from each region, ensuring a consistent extraction between both spectra.

Concerning the individual performance of each comparator, SIFT exhibits very low error rates at the original image size, but this comparator is computationally heavy both in processing times and template size. In this paper, we use the SIFT detector with the same parametrization employed in [75] for iris images of size 640 × 480. In the work [75], the iris region represented ∼1/8 of the image only, leading to some hundreds of key-points per image. However, images of the Cross-Eyed database are of 613 × 701 pixels, and the periocular ROI occupies a considerably bigger area than the iris region, leading to an average of ∼1900 key-points per image (annular ROI) or ∼2543 (full ROI). To match two images, it is needed to compare each key-point of one image against all key-points of the other image to find a pair match. This increases the computation time exponentially when the number of key-points per image increases, which is one of the drawbacks of key-point based comparators [1]. The other comparators employed have templates of fixed size, thus comparison is made very efficiently using distance measures involving a number of fixed calculations. It can be also seen that the use of annular (AR) or full ROI (FR) does not produce a significant difference with SIFT. This suggests that the annular ROI is sufficient, and the key-points of the corner areas incorporated with the full image (Fig. 3) do not contribute to a better performance with the Cross-Eyed database, while the comparison time is increased by 62% (Table 4). Therefore, we carry forward the AR configuration of SIFT to the fusion experiments of the next section. On the other hand, if the size of the input image is reduced to match the CNNs (224 × 224), the lower amount of detected key-points (only 92 on average) produces that both same-sensor and cross-spectral performance degrades one of two orders of magnitude. When this happens, SIFT becomes worse than

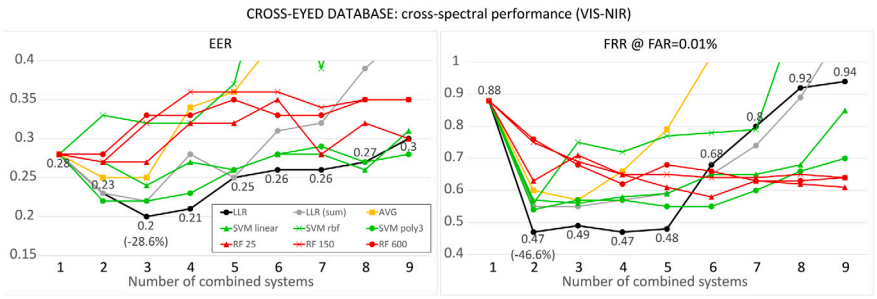

Fig. 10. Cross-Eyed database, test set: Verification results for an increasing number of fused comparators. Best seen in colour. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

e.g. DenseNet201 or ResNet 101, and comparable to VGG-Face in some DET regions (Fig. 9). This can also serve as indication of the strength of the CNNs, which match SIFT's performance if the image size of the latter is reduced to be equal, and they also rank ahead of the other comparators while using a smaller input image size.

In general, there is an inverse proportion between the error rates and the template size. The comparators with the best performance (SIFT, NTNU and the three CNNs) are also the ones with the biggest feature vector (see Table 3). It is remarkable the performance of NTNU as well, surpassing the CNNs in some cases, but with a smaller feature vector. When it comes to cross-spectral comparisons, however, the CNNs provide better performance. This is observed especially with the deeper networks (ResNet101 and DenseNet201), highlighting the capability of these powerful descriptors pre-trained with millions of images. In the DET curves of Fig. 9, it can be better appreciated the superiority of the three CNNs for cross-spectral comparisons w.r.t. the other comparators (apart from SIFT). It is also remarkable the behaviour of DenseNet201, which provides the second-best result of all comparators, but with a feature vector much smaller than the other CNNs. Among the three CNNs used in this paper, DenseNet201 is the one providing the best performance on the original task for which they were trained (ImageNet), so it could be expected that such superiority is transferred to other tasks as well. It is also worth noting the relatively good cross-spectral EER values of some *light* comparators such as LBP or HOG. With a feature vector of only 384 real numbers and an EER of 5%–6%, they would enable low-security applications where computational resources are limited.

5.4. Results: Fusion of periocular comparators

We then carry out fusion experiments using all the available comparators, according to the fusion schemes presented in Section 4. We have tested all the possible fusion combinations. Whenever training is needed (i.e. to compute calibration weights, z-normalization, SVM, or Random Forest models), the training set of the Cross-Eyed database is used. In Fig. 10, we show the best results obtained for an increasing number M of combined comparators. Following the protocol of Cross-Eyed 2016, the best combinations are chosen based on the lowest cross-spectral FRR @ FAR = 0.01%. Then, the corresponding EER of the chosen combinations is reported as well in Fig. 10. We use the two mentioned calibration possibilities of the fusion method (Fig. 5): (BX-MATH[148]) the scores from all comparators are calibrated together ($N = M$ in Eq. (2)), or (ii) the score of each comparator is calibrated separately ($N = 1$) and the resulting calibrated scores are summed. These cases are shown in Fig. 10 as 'LLR' and 'LLR (sum)', respectively.

As it can be observed, a substantial performance improvement can be obtained when combining several comparators. The best cross-spectral performance is obtained with a combination of 2 to 3 comparators. The FRR remains approximately constant until 5 comparators are combined, and then it deteriorates when including more. The EER, nevertheless, deteriorates earlier. We also observe that the probabilistic fusion method based on calibration (LLR) outperform all the others, demonstrating its superiority. This is more evident at low FAR, with a relative FRR reduction of ∼47% in comparison to using one comparator only. It is also better if all scores are calibrated together, rather than calibrating them individually and then summing them up ('LLR' vs.'LLR (sum)'). Regarding the other fusion methods, the SVM with a linear or polynomial kernel stands out in comparison to the others. The polynomial kernel shows equal or better performance in some cases, but such kernel is much slower to train. It is also worth noting that the simple average rule (AVG) provides similar performance than trained approaches like the SVM, although it deteriorates quickly with the combination of more than 3 comparators. On the other hand, the Random Forest approach performs among the worst, regardless of the number of decision trees employed.

In Table 6, we show the comparators involved in the best fusion cases. For the sake of space, we only provide results with a selection of fusion approaches, according to the observations made above when discussing Fig. 10: the LLR method (best case), SVM linear (a good runner-up which is also faster to train than its polynomial counterpart), and AVG or AVERAGE (a simple approach that does not need training). To allow a more comprehensive analysis, we also provide not only the best cases but also the second and third best combinations for a given number of comparators. It can be seen that the best combinations for any given number of comparators always involve the SIFT method. The excellent accuracy of the SIFT comparator is not jeopardized by the fusion with other comparators that have a performance one or two orders of magnitude worse, but it is complemented to obtain even better cross-spectral error rates, especially with trained approaches. A careful look at the combinations of Table 6 shows that the CNN comparators are also chosen first for the fusion. Together with SIFT, they are the comparators with the best individual performance, and they appear to be very complementary too. However, it should not be taken as a general statement that the best fusion combination always involves the best individual comparators. Different fusion algorithms may lead to different results [86,94]. For example, the best FRR with the simple average rule involves the SAFE comparator. It is also worth noting that other comparators with worse individual performance and not based on deep networks (such as SAFE, LBP, or NTNU) are also selected in combinations that have a performance nearly as good as the best cases. At the same time, this shows the power of the fusion approaches employed, and especially of the calibration method, which are capable of reducing error rates substantially by fusion of comparators with very heterogeneous performance and different feature representations.

To further illustrate the benefit of using calibrated scores, we plot in Fig. 11 the False Acceptance/False Rejection (FA/FR) curves of the individual systems. This is done using raw scores of each system (left), normalized scores using z-score normalization (centre), and calibrated scores (right). One selected fusion case of Table 6 (best combination of three systems: SIFT+LBP+ResNet101) is also plotted using average of normalized scores (centre) and score calibration (right). It can be seen that the raw scores of each system lies in a different range, even if all comparators are expected to produce a score between 0 and 1

Table 6
Cross-Eyed database, test set: Verification results for an increasing number of fused comparators. The best combinations are chosen based on the lowest FRR @ FAR = 0.01% of cross-spectral experiments. The best result of each column is marked in bold.

Cross-Eyed database: cross-spectral performance (VIS–NIR)																														
	LLR fusion											Average fusion											SVM linear fusion									
# comparators	safe	gabor	sift	lbp	hog	ntnu	vgg-face	resnet101	densenet201	EER (%)	FRR (%)	safe	gabor	sift	lbp	hog	ntnu	vgg-face	resnet101	densenet201	EER (%)	FRR (%)	safe	gabor	sift	lbp	hog	ntnu	vgg-face	resnet101	densenet201	EER (%)	FRR (%)
1			x							0.28	0.88			x							0.28	0.88			x							0.28	0.88
									x	2.09	10.09									x	2.09	10.09									x	2.09	10.09
								x		2.62	12.51								x		2.62	12.51								x		2.62	12.51
2			x						x	0.23	**0.47**			x					x		**0.25**	0.6			x				x			0.27	0.57
			x					x		0.21	0.48			x						x	**0.25**	0.62			x						x	0.24	0.59
	x		x							0.26	0.52			x				x			0.33	0.66			x					x		0.23	0.61
3			x	x				x		**0.2**	0.49	x		x					x		**0.25**	**0.57**			x					x	x	0.24	**0.56**
			x					x	x	0.21	0.49			x			x			x	0.28	0.67			x				x	x		0.27	0.58
	x		x				x			0.29	0.5	x		x						x	0.28	0.68			x	x			x			0.27	0.6
4			x	x				x	x	0.21	**0.47**	x		x	x				x		0.34	0.66			x				x	x	x	0.27	0.58
			x	x			x	x		**0.2**	0.5	x		x			x		x		0.3	0.69			x	x			x	x		0.27	0.59
	x		x				x	x		0.31	0.51	x		x			x			x	0.28	0.72			x	x				x	x	0.25	0.6
5	x		x	x			x	x		0.25	0.48	x		x	x		x		x		0.36	0.79			x	x			x	x	x	0.26	0.59
	x		x	x				x	x	0.32	0.59	x		x		x	x		x		0.34	0.88			x	x	x			x	x	**0.22**	0.63
		x	x	x				x	x	0.25	0.64	x		x	x		x			x	0.34	0.88	x		x	x				x	x	**0.22**	0.64
6	x	x	x	x	x		x			0.26	0.68	x		x	x	x	x		x		0.43	1.02			x	x		x	x	x	x	0.28	0.65
		x	x	x	x		x	x		0.27	0.69	x		x	x		x		x	x	0.43	1.04	x		x			x	x	x	x	0.28	0.65
		x	x	x	x			x	x	0.25	0.7	x		x	x	x	x			x	0.41	1.07			x	x	x	x	x	x		0.27	0.66
7	x		x	x		x	x	x	x	0.26	0.8	x		x	x	x	x		x	x	0.51	1.11	x		x	x		x	x	x	x	0.28	0.65
	x	x	x	x	x		x	x		0.29	0.81	x		x	x	x	x	x		x	0.55	1.26	x		x	x	x		x	x	x	0.24	0.67
			x	x	x	x	x	x	x	0.22	0.83	x	x	x	x	x	x			x	0.68	1.32	x		x	x	x	x	x	x		0.27	0.67
8	x		x	x	x	x	x	x	x	0.27	0.92	x	x	x	x	x	x		x	x	0.74	1.49	x		x	x	x	x	x	x	x	0.26	0.68
	x	x	x	x	x	x		x	x	0.31	0.93	x		x	x	x	x	x	x	x	0.64	1.51	x	x	x	x	x		x	x	x	0.26	0.81
	x	x	x	x	x	x	x		x	0.29	0.94	x	x	x	x	x	x	x	x		0.81	1.62	x	x	x	x	x	x	x	x		0.31	0.84
9	x	x	x	x	x	x	x	x	x	0.3	0.94	x	x	x	x	x	x	x	x	x	0.84	1.96	x	x	x	x	x	x	x	x	x	0.31	0.85

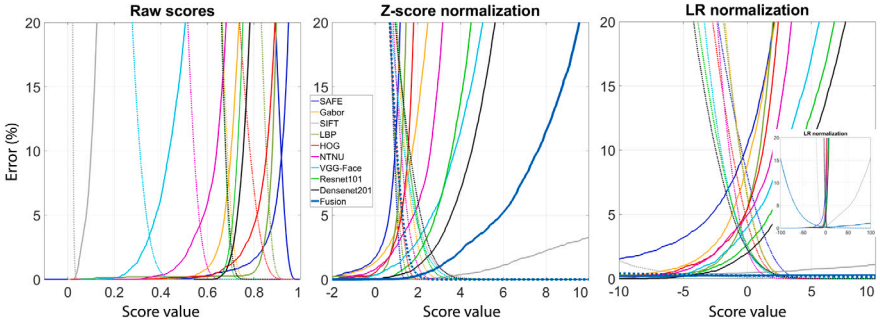

Fig. 11. Cross-Eyed database, test set: cross-spectral FA/FR curves of the individual systems (left: with raw scores, middle: after z-score normalization, right: after mapping to log-likelihood ratios). Solid curves represent FR curves, and dashed curves represent FA curves. The 'fusion' curves on the centre and right plots represent the fusion of SIFT+LBP+Resnet101 (see the main text for details). Best seen in colour and zoomed. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

([−1, 1] with SAFE). After z-score normalization, the impostor score distributions become aligned to a certain degree, since such normalization converts them to zero mean and unit variance. Also, the extent to which the genuine distributions spread are indicative of the performance of each system (in order: SIFT (grey), DenseNet201 (black), ResNet101 (green), etc.). However, this cannot always be expected, since the fusion (blue thick curve) is situated between the curves of the individual systems involved due to scores being averaged. The EER of each system occurs as a different score value too. Similar effects can be expected with other popular normalization techniques like max–min, tanh, etc. [24]. When scores are normalized by calibration, two phenomena occur: (*i*) the FA and FR curves cross at ∼0 score (the EER is always situated at this point), since a positive log-likelihood-ratio output supports the genuine (mated) decision, and a negative value the opposite; and (*ii*) the spread and order of the curves are indicative of the performance of each system. For example, the SIFT curves (grey) have a smaller slope and reach higher log-likelihood-ratios (both positive and negative), due to this system being significantly better than the others (Table 5). The FA and FR curves of the other systems are then ordered (both in positive and negative sides): DenseNet201 (black), ResNet101 (green), VGG-Face (blue), etc. Furthermore, after the fusion (blue thick curves), the slope of the curves is even less, reaching even higher score values on both extremes. Given that the performance of the fusion is better than any of the other systems, both the genuine

Table 7
Comparison with results of the Cross-Eyed 2016 Competition [27]. GF2 is the Generalized FRR (GFRR) at a Generalized FAR (GFAR) of 0.01%. The GFRR and GFAR are generalizations of the FRR and FAR to include Failure to Acquire (FTA) and Failure to Enrol (FTE) rates, according to ISO/IEC standards [95]. The ranking in the evaluation of the submitted approaches is also given. For more information, refer to [27].

Cross-Eyed database: Cross-spectral performance (VIS–NIR)												
Approach	safe	gabor	sift	lbp	hog	Training set		Test set		Competition [27]		
						EER	FRR	EER	FRR	EER	GF2	Rank
HH3		x		x	x	4.5	16.77	4.86	24.59	6.02	11.42	3rd
HH2	x	x		x	x	3.02	12.63	4.51	19.75	5.24	9.14	2nd
HH1	x	x	x	x	x	0	0	0.28	0.83	0.29	0	1st

and impostor scores are pushed towards the extremes of the horizontal axis. This reflects the probabilistic meaning of calibrated scores, in the sense that a better performance translates to a reduced uncertainty via higher absolute score values.

5.5. Results: Comparison with the Cross-Eyed 2016 competition

Table 7 shows the results of the submission of Halmstad University to the Cross-Eyed 2016 competition. We provide both the results reported by the organizers [27], and our own computations on the training and test sets of the database using the executables submitted and the protocol described in Section 5.1. For the evaluation, only the SAFE, GABOR, SIFT, LBP, and HOG comparators were available. We contributed with three different fusion combinations, named HH1, HH2, and HH3, with the HH1 combination obtaining the first position in the competition. Two key differences in the results reported in Table 7 in comparison with the present paper are that in our executables: (*i*) the score of each comparator was calibrated separately, and the resulting calibrated scores were summed up; and (*ii*) the LBP and HOG comparators employed the Euclidean distance (which is the popular choice in the literature with these methods, instead of χ^2). At the time of submission, the test set had not been released, so our decisions could only be based on the results on the training set. We observed that the SIFT comparator already provided cross-spectral error rates of nearly 0% on the training set (not shown in Table 7). However, it was reasonable to expect a higher error with a bigger dataset, as demonstrated later when the test set was released. Therefore, we contributed to the competition with a fusion of the five comparators available (called HH1) to be able to better cope with the generalization issue that is expected when performance is measured in a bigger set of images. Indeed, in Table 7 it can be seen that performance on the test set is systematically worse than on the training set. Since the combination of the five available comparators is computationally heavy in template size (due to the SIFT comparator), we also contributed by removing SIFT (combination HH2), and by further removing SAFE (combination HH3), which has a feature extraction time considerably higher than the rest of the comparators in our implementation (see Table 4). Thus, our motivation behind HH2 and HH3 was to reduce template size and feature extraction time. Some differences are observable between our results with the test set and the results reported by the competition [27]. We attribute this to two factors: (*i*) the additional 10 subjects included in the test set released, which were not used during the competition, and (*ii*) the employment of a different test protocol since it is not specified by the organizers the exact images used for impostor trials during the competition. Therefore, the experimental framework used in this paper is not exactly the same employed in the Cross-Eyed competition.

6. Cross-sensor (VIS-VIS) smartphone periocular recognition

6.1. Database and protocol

In the cross-sensor experiments of this section, we use the Visible Spectrum Smartphone Iris (VSSIRIS) database [28], which has images from 28 subjects (56 eyes) captured using the rear camera of two smartphones (Apple iPhone 5S, of 3264 × 2448 pixels, and Nokia Lumia 1020, of 3072 × 1728 pixels). They have been obtained in unconstrained conditions under mixed illumination (natural sunlight and artificial room light). Each eye has 5 samples per smartphone, thus 5 × 56 = 280 images per device (560 in total). The acquisition is made without flash, in a single session and with semi-cooperative subjects. Fig. 7 (bottom) shows some examples.

Table 8
VSSIRIS database: Experimental protocol.

VSSIRIS database		
Protocol (28 subjects)	Same-sensor	Cross-sensor
Genuine	56 × (4 + 3 + 2 + 1) = 560	56 × 5 × 5 = 1400
Impostor	56 × 55 = 3080	56 × 55 = 3080

All images of VSSIRIS are annotated manually, so the radius and centre of the pupil and sclera circles are available. Images are resized via bicubic interpolation to have the same sclera radius (set to R_s = 145, the average radius of the whole database). We use the sclera for normalization since it is not affected by dilation. Then, images are aligned by extracting a square region of $6R_s \times 6R_s$ (871 × 871) around the sclera centre. This size is set empirically to ensure that all available images have sufficient margin to the four sides of the sclera centre. Here, there is sufficient availability to the four sides of the eye, so the normalized images have the eye centred in the image, as can be seen in Fig. 3 (bottom). Images are further processed by Contrast-Limited Adaptive Histogram Equalization (CLAHE) [38] to compensate for variability in local illumination.

We carry out verification experiments, with each eye considered a different user. We compare images both from the same device (*same-sensor*) and from different devices (*cross-sensor*). Genuine trials are obtained by comparing each image of an eye to the remaining images of the same eye, avoiding symmetric comparisons. Impostor trials are done by comparing the 1st image of an eye to the 2nd image of the remaining eyes. The experimental protocol is summarized in Table 8. The smaller size of VSSIRIS in comparison with the Cross-Eyed database results in the availability of fewer scores. Therefore, whenever a parameter needs training, 2-fold cross-validation [96] is used, dividing the available number of users in two partitions. Otherwise, we report results employing the entire VSSIRIS database.

The parameters of the periocular comparators are as follows. As with the Cross-Eyed database, they are designed to adapt dynamically to the size of the image, being the sclera boundary the only necessary input. Regarding the SAFE comparator, the annular band of the first circular ring starts at the sclera circle (R = 145 pixels), and the band of the last ring ends at the boundary of the image, resulting in a ROI of 871 × 871 pixels around the eye centre. The availability of sufficient margin around the four sides of the eye makes possible to have a bigger ROI with VSSIRIS, as can be shown in Fig. 3, third column. This availability also allows one extra row in the grid employed with GABOR, LBP and HOG comparators, having 8 × 8 = 64 non-overlapping blocks. Given the size of the input image, each block has 109 × 109

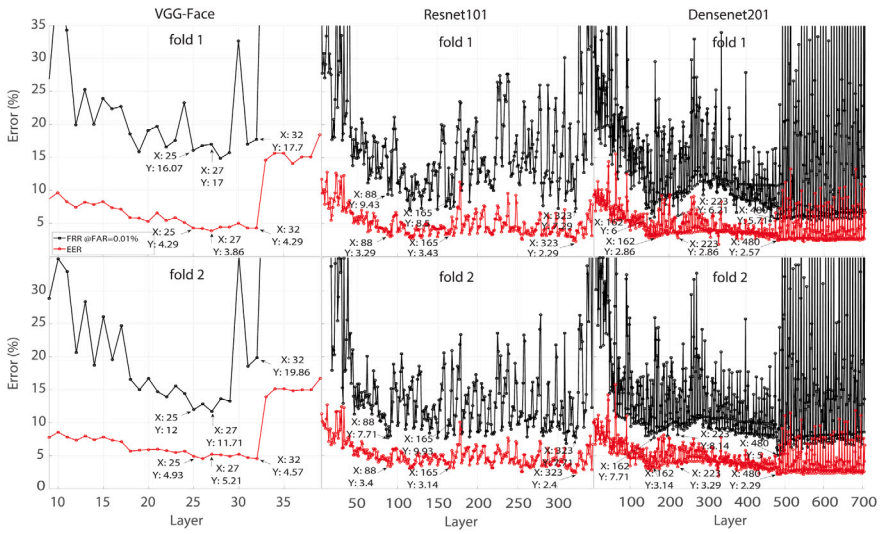

Fig. 12. VSSIRIS database: Cross-sensor accuracy (VIS–VIS) of different CNN layers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

pixels. For consistency with Cross-Eyed, the eight blocks of the image centre are not considered, effectively resulting in 56 blocks (some more than Cross-Eyed, which has 48 blocks of size 88 × 88 each). The GABOR comparator employs filter wavelengths spanning from 55 to 7 pixels, which are set proportional to the block size as 109/2≈55 to 109/16≈7. Regarding VGG-Face, Resnet101 and Densenet201, images are resized to 224 × 224, which are the input dimensions of these CNNs. With SIFT, we keep as baseline the use of the annular ROI, but for comparison purposes, we also evaluate the entire input image (both at the original size of 871 × 871 and at 224 × 224). Table 3 (third column) indicates the size of the feature vector for a given periocular image with the different comparators employed. The full ROI produces an average of 3824 SIFT key-points (489 472 values), and ∼3000 with the annular ROI (384 000 values), which are higher values than Cross-Eyed, since the image is bigger. At 224 × 224, there are 130 key-points per image on average (16 512 values). Experiments have been done in the same machine and with the same algorithm implementations than Cross-Eyed (Section 5.1). The feature extraction and comparison times are given in Table 4 (right).

6.2. Results: Finding the optimum layer of the convolutional Neural Networks

We first identify the optimum layer of each CNN. The cross-sensor accuracy of each network is given in Fig. 12 for each cross-validation fold. When selecting the best layer, we have tried to find the one that gives optimum performance both with the Cross-Eyed and the VSSIRIS databases simultaneously. However, it has not always been possible. According to the discussion in Section 5.2, the best layers with VSSIRIS are layer 25 (VGG-Face), layer 323 (ResNet101), and layer 480 (DenseNet201). It can be seen as well that the optimum layers of VSSIRIS are the same for the two folds. With Cross-Eyed, on the other hand, the best layers were not so deep: 165 (ResNet101) and layer 223 (DenseNet201).

6.3. Results: Individual comparators

The performance of individual comparators is then reported in Table 9. Similarly as Section 5, we adopt as measures of accuracy the EER and the FRR at FAR = 0.01%. In Fig. 13, we give the DET curves of the cross-sensor experiments.

By comparing Tables 5 and 9, it can be observed that same-sensor experiments with the VSSIRIS database usually exhibit lower error

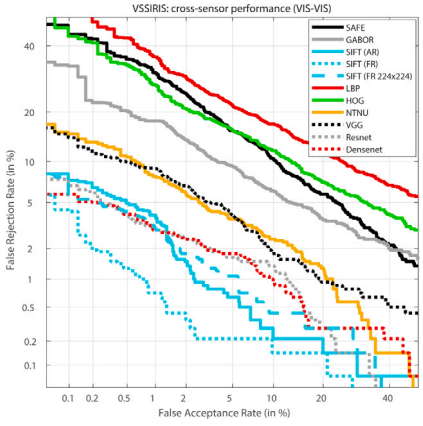

Fig. 13. VSSIRIS database: Verification results of the individual comparators. Best seen in colour. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

rates for any given comparator. Possible explanations might be that the ROI of VSSIRIS images is bigger (871 × 871 vs.613 × 701), or that the VSSIRIS database has fewer users (28 vs.90 subjects). On the opposite side, cross-sensor error rates with VSSIRIS are significantly worse for some comparators (e.g. SIFT, HOG, NTNU, or VGG-Face). *Lighter* comparators such as LBP or HOG are not capable of providing good cross-sensor performance in low-security applications either (EER of 11% or higher). The difference is especially relevant with the SIFT comparator, where cross-sensor error rates on Cross-Eyed (Table 5) were 0.28% (EER) and 0.88% (FRR), but here they increase one order of magnitude, up to 1.6% (EER) and 12.7% (FRR) (annular ROI). This is despite the higher number of SIFT key-points per image with VSSIRIS due to higher image size (∼3000 vs.∼1900 on average). It is thus interesting that the comparators employed in this paper are more robust to the variability between images in different spectra (NIR and VIS) than the variability between images in the same (VIS) spectrum captured with two different smartphones. Such effect can also be seen in that the SIFT comparator is more sensitive to changes in the ROI with VSSIRIS. With the full ROI (FR), cross-sensor errors are divided by two, so a bigger ROI can be seen as a way to counteract cross-sensor variability in this case. Another difference here is that if the image size is reduced to 224 × 224, SIFT does not degrade as much, being in some

Table 9
VSSIRIS database: Verification results of the individual comparators. The relative variation of cross-sensor performance with respect to the best same-sensor performance is given in brackets (for the SIFT comparator, the result is calculated w.r.t. the Nokia performance, since the iPhone performance is 0%, which would result in Inf due to division by zero; if both Nokia and iPhone performance is 0%, no value is given). AR = annular ROI. FR = Full ROI.

Comparator	Equal Error Rate (EER)			FRR @ FAR = 0.01%		
	Same sensor			Same sensor		
	iPhone	Nokia	Cross-sensor	iPhone	Nokia	Cross-sensor
SAFE	1.6%	2.6%	10.2% (+537.5%)	4.6%	11.1%	50.9% (+1006.5%)
GABOR	2.1%	1.5%	7.3% (+386.7%)	4.3%	8.9%	39.1% (+809.3%)
SIFT (AR 871 × 871)	0%	0.1%	1.6% (>1500%)	0%	0.7%	12.7% (>1700%)
SIFT (FR 871 × 871)	0%	0%	0.82% (–)	0%	0%	6.25% (–)
SIFT (FR 224 × 224)	0%	0%	1.79% (–)	0%	0%	10.54% (–)
LBP	4.8%	4.9%	14.1% (+193.8%)	6.8%	16.8%	71.2% (+947.1%)
HOG	3.9%	4.5%	11% (+182.1%)	5.2%	17.3%	70.7% (+1259.6%)
NTNU	0.7%	0.7%	4.1% (+480%)	0.9%	1.8%	23.1% (+2500%)
VGG-Face	0.9%	0.7%	4.4% (+528.6%)	1.6%	1.3%	20.8% (+1500%)
Resnet101	0.5%	0%	2.3% (–)	0.7%	0.4%	10.3% (+2475%)
Densenet201	0.5%	0%	2.4% (–)	0.7%	0.2%	6.2% (+3000%)

regions of the DET at the same level than the baseline annular ROI (AR) or than some CNNs. It should be noted, though, that images in Cross-Eyed are obtained with a dual spectrum sensor, which captures NIR and VIS images synchronously. Thus, in practice, there is no scale, 3D rotation or time-lapse difference between corresponding NIR and VIS samples. Only a spatial offset between the two exist in the plane perpendicular to the optical axes of the cameras due to the sensors not being perfectly calibrated (which can be noticed in Fig. 7), so images are expected to be very well aligned after cropping. This synchronicity and absence of time span could be one of the reasons of the better cross-spectral performance obtained with the Cross-Eyed database, or the less sensitivity of the SIFT method to changes in the ROI.

Another observation is that same-sensor performance with VSSIRIS is sometimes very different depending on the smartphone employed, even if they involve the same subjects and images are resized to the same size. Contrarily, same-sensor performance with Cross-Eyed tends to be similar regardless of the spectrum employed (Table 5), which might be explained as well by the synchronicity in the acquisition mentioned above. Previous works have suggested that discrepancy in colours between VIS sensors can lead to variability in performance, which is further amplified when images from such sensors are compared among them. The sensitivity of SIFT to changes in the ROI can also be an indicative of this. Although we apply local adaptive contrast equalization, our results suggest that other device-dependent colour correction might be of help [45]. Another difference observed here is that the best individual comparator (in terms of FRR) is not SIFT. With Cross-Eyed, SIFT was the best by a large margin, but here, other comparators have similar or better performance (e.g. DenseNet201, ResNet101). This is despite the higher number of SIFT key-points per image with VSSIRIS mentioned above. Nevertheless, the correlation between bigger template size and lower error rates remains since the comparators with the best performance (SIFT, NTNU and the three CNNs) are also the ones with the biggest feature vector. The superiority of these comparators can also be observed in the DET curves of Fig. 13.

6.4. Results: Fusion of periocular comparators

We now carry out fusion experiments using all the available comparators. Whenever a fusion method needs training, 2-fold cross-validation [96] was used, dividing the available number of users in two partitions. We have also tested here all the possible fusion combinations, with the best combinations chosen based on the lowest cross-sensor FRR @ FAR = 0.01%. The best results obtained for an increasing number M of combined comparators is given in Fig. 14 (average values of the two folds). The comparators involved in the best fusion cases are also given in Table 10 (as in Section 5.4, the table only shows the results of a selection of fusion approaches).

Similarly as Cross-Eyed, cross-sensor performance is also improved significantly here by fusion. The relative EER and FRR improvement of the best fusion case is even bigger, being 87.5% and 95.2%, respectively. This is high in comparison with the reductions observed with Cross-Eyed, which were in the order of 30%–40%. It is also remarkable that similar or even better absolute performance values are obtained with VSSIRIS. This is despite the worse performance observed in the individual comparators, as discussed in the previous section. However, it comes at the price of needing more comparators to achieve maximum performance. Even if the biggest performance improvement also occurs after the fusion of two or three comparators, the smallest error is obtained with the fusion of four comparators. In contraposition, Cross-Eyed needed only two or three (see Fig. 10).

The fusion methods evaluated also rank in the same order here (see Fig. 14). The probabilistic fusion method based on calibration (LLR) outperforms all the others, followed by SVM linear and polynomial. The simple average rule also matches the performance of other trained approaches in some points, but it deteriorates quickly as more comparators are combined. Lastly, the Random Forest approach performs the worst in general. In addition, the SIFT comparator is also decisive to achieve lower error rates, as it is always selected in any combination (Table 10). The CNN comparators are also selected first, but to achieve the best performance, the role of other comparators are decisive with this database. The best FRR, for example, is given by the combination of SAFE, SIFT, LBP and DenseNet201. The same can be said with other fusion methods. The best FRR with the average fusion involves SIFT, NTNU, and DenseNet201, while the best FRR with the linear SVM engages SAFE, GABOR, SIFT, HOG and DenseNet201.

Fig. 15 provides the FA/FR curves of the systems with different score normalizations. A selected fusion case is also plotted (SAFE+SIFT+LBP+DenseNet201, best combination of four systems in Table 10). The same observations than Section 5.4 can be made, in the sense that calibration provides alignment of genuine and impostor distribution around zero, and that the arrangement and spread of the distributions to both sides of the horizontal axis are indicative of the relative performance among systems.

7. Conclusion

Periocular biometrics has rapidly evolved to competing with face or iris recognition [1,2]. The periocular region has shown to be as discriminative as the full face, with the advantage that it is more tolerant to variability in expression, blur, downsampling [97], or occlusions [12,98]. Under difficult conditions, such as people walking by acquisition portals, [99–101], distant acquisition, [102,103], smartphones, [45], webcams, or digital cameras, [33,91], the periocular modality is also

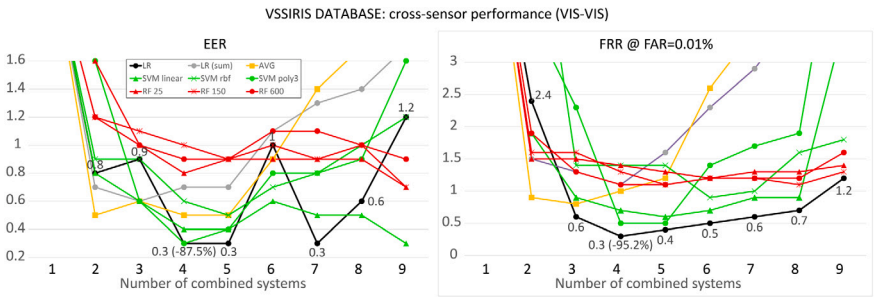

Fig. 14. VSSIRIS database, test set: Verification results for an increasing number of fused comparators. Best seen in colour. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 10
VSSIRIS database: Verification results for an increasing number of fused comparators. The best combinations are chosen based on the lowest FRR @ FAR = 0.01% of cross-sensor experiments. The best result of each column is marked in bold.

VSSIRIS DATABASE: cross-sensor performance (VIS–VIS)																													
	LLR FUSION											AVERAGE FUSION											SVM LINEAR FUSION										
# comparator	safe	gabor	sift	lbp	hog	ntnu	vgg-face	resnet101	densenet201	EER(%)	FRR (%)	safe	gabor	sift	lbp	hog	ntnu	vgg-face	resnet101	densenet201	EER(%)	FRR (%)	safe	gabor	sift	lbp	hog	ntnu	vgg-face	resnet101	densenet201	EER (%)	FRR (%)
1									x	2.4	6.2									x	2.4	6.2									x	2.4	6.2
								x		2.3	10.3								x		2.3	10.3								x		2.3	10.3
			x							1.6	12.7			x							1.6	12.7			x							1.6	12.7
2			x						x	0.8	2.4			x						x	**0.5**	0.9			x					x		0.8	1.9
			x					x		0.9	2.4			x					x		0.6	1.6			x						x	0.7	1.9
			x				x			1	2.8			x				x			0.9	2.8			x				x			0.9	2.7
3	x		x						x	0.9	0.6			x			x			x	0.6	**0.8**	x		x						x	0.6	0.9
			x			x			x	1.4	0.6	x		x						x	**0.5**	0.9		x	x						x	0.6	1
			x	x					x	**0.3**	0.7			x					x	x	**0.5**	1.1			x	x				x		0.5	1
4	x		x	x					x	**0.3**	**0.3**	x		x			x			x	**0.5**	1	x	x	x						x	0.4	0.7
	x		x				x	x		0.6	0.5	x		x			x		x		0.7	1	x		x		x				x	0.6	0.8
	x	x	x						x	0.5	0.5	x		x					x	x	**0.5**	1.1	x		x	x					x	**0.2**	0.8
5	x		x	x				x	x	**0.3**	0.4	x		x			x		x	x	**0.5**	1.2	x	x	x		x				x	0.4	**0.6**
	x		x				x	x	x	0.5	0.4	x		x				x	x	x	0.9	2.3	x		x	x				x	x	0.3	0.7
			x			x	x	x	x	1.4	0.6	x		x	x				x	x	1.1	2.4	x		x		x	x			x	0.6	0.8
6	x		x			x	x	x	x	1	0.5	x		x			x	x	x	x	0.9	2.6	x		x		x	x		x	x	0.6	0.7
	x		x		x		x	x	x	**0.3**	0.5	x		x		x	x		x	x	1	2.7	x	x	x		x			x	x	0.5	0.8
	x		x	x			x	x	x	**0.3**	0.5	x		x	x		x		x	x	1	3	x	x	x		x		x		x	0.6	0.8
7	x	x	x	x	x		x		x	**0.3**	0.6	x	x	x		x	x		x	x	1.4	3.5	x	x	x			x	x	x	x	0.5	0.9
	x	x	x	x	x	x		x		1	0.6	x		x		x	x	x	x	x	1.4	3.6	x	x	x		x		x	x	x	0.6	0.9
	x	x	x	x			x	x	x	**0.3**	0.7	x		x	x	x	x		x	x	1.2	3.8	x	x	x		x	x		x	x	0.6	0.9
8	x	x	x		x	x	x	x	x	0.6	0.7	x	x	x		x	x	x	x	x	1.7	4.1	x	x	x		x	x	x	x	x	0.5	0.9
	x	x	x	x	x	x	x	x		1.1	0.8	x	x	x	x	x	x		x	x	1.8	4.2	x	x	x	x	x		x	x	x	**0.2**	1.8
	x	x	x	x	x	x		x	x	0.9	0.9	x		x	x	x	x	x	x	x	1.6	4.5	x	x	x	x	x	x	x		x	0.3	2.4
9	x	x	x	x	x	x	x	x	x	1.2	1.2	x	x	x	x	x	x	x	x	x	1.9	4.9	x	x	x	x	x	x	x	x	x	0.3	3.6

shown to be clearly superior to the iris modality, mostly due to the small size of the iris or the use of visible illumination. The COVID-19 pandemic has also imposed the necessity of developing technologies capable of dealing with faces occluded by protective face masks, often with just the periocular area visible [7–9].

As biometric technologies are extensively deployed, it will be common to compare data captured with different sensors or from uncontrolled non-homogeneous environments. Unfortunately, the comparison of heterogeneous biometric data for recognition purposes is known to decrease performance significantly [11]. Hence, as new practical applications evolve, new challenges arise, as well as the need for developing new algorithms to address them. In this context, we address in this paper the problem of biometric sensor interoperability, with recognition by periocular images as test-bed.

Inspired by our submission to the 1st Cross-Spectral Iris/Periocular Competition (Cross-Eyed) [27], we propose to mitigate such problem via a multialgorithm fusion strategy at the score level that combines up to nine different periocular comparators. The aim of this competition was to evaluate periocular recognition algorithms when images from visible and near-infrared spectra are compared. We follow a probabilistic score fusion approach based on linear logistic regression [41,81]. With this method, scores from multiple comparators are fused together not only to improve the discriminating ability but also to produce log-likelihood ratios as output scores. This way, output scores are always in a comparable probabilistic domain since log-likelihood ratios can be interpreted as a degree of support to the *target* or *non-target* hypotheses. This allows the use of Bayes thresholds for optimal decision-making, avoiding the need to compute comparator-specific thresholds. This is

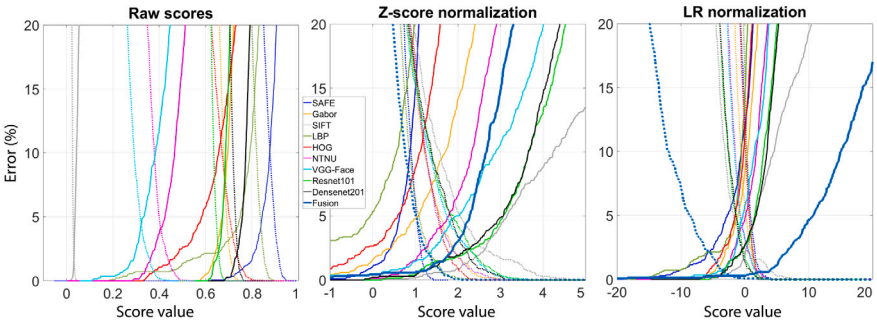

Fig. 15. VSSIRIS database: cross-sensor FA/FR curves of the individual systems (left: with raw scores, middle: after z-score normalization, right: after mapping to log-likelihood ratios). Solid curves represent FR curves, while dashed curves represent FA curves. The 'fusion' curves on the centre and right plots represent the fusion of SAFE+SIFT+LBP+Densenet201 (see the main text for details). Best seen in colour and zoomed. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

essential in operational conditions since the threshold is critical to determine the accuracy of the authentication process in many applications. In the experiments of this paper, this method is shown to surpass other fusion approaches such as the simple arithmetic average of normalized scores [24] or trained algorithms such as Support Vector Machines [25] or Random Forest [26]. This employed fusion approach has been applied previously to cross-sensor comparison of face or fingerprint modalities [23] as well, also providing excellent results in other competition benchmarks involving these modalities [43]. We employ in this paper three different comparators based on the most widely used features in periocular research [12], as well as three in-house comparators that we proposed recently [32–34], and three comparators based on deep Convolutional Neural Networks [35–37]. The proposed fusion method, with a subset of the periocular comparators employed here, was used in our submission to the mentioned Cross-Eyed evaluation, obtained the first position in the ranking of participants. This paper is complemented with cross-sensor periocular experiments using images from the same spectrum as well. For this purpose, we use the Visible Spectrum Smartphone Iris database (VSSIRIS) [28], which contains images in the visible range from two different smartphones.

We first analyse the individual comparators employed not only from the point of view of its cross-sensor performance (Figs. 9 and 13), but also taking into account its template size and computation times (Tables 3 and 4). We observe that the comparator having the biggest template size and computation time is usually the most accurate in terms of individual performance, also contributing decisively to the fusion. In the experiments reported in this paper, significant improvements in performance are obtained with the proposed fusion approach, leading to an EER of 0.2% in visible-to-near-infrared comparisons (Fig. 10) and 0.3% in visible-to-visible comparison of smartphone images (Fig. 14). The FRR in high-security environments (at FAR = 0.01%) is also very good, being 0.47% and 0.3%, respectively.

Interestingly, the best performance is not obtained necessarily by the combination of all available comparators. Instead, the best results are obtained by fusion of just two to four comparators. A fundamental problem in classifier combination is to determine which systems to retain in order to attain the best results [104]. The systems retained are not necessarily the best individual ones, especially if they are not sufficiently complementary (for example, if they employ similar features) [86]. When the comparators are properly chosen (in our case, found by exhaustive search), the performance increases quickly with the addition of a small number of them. Then, it tends to stabilize until the addition of new ones actually decreases the performance. The need to retain the best features only, and the mentioned performance 'peaking' effect, is well documented [104], and it can be attributed to the correlation between classifiers or to the effect of a limited sample size. Such phenomenon have been also observed in other related studies in biometrics [86,91,105–107]. It is also worth noting that the comparators producing the best fusion performance (Tables 6 and 10) have an individual performance that differs in one or two orders of magnitude in some cases. In the probabilistic approach employed, each comparator is implicitly weighted by its individual accuracy, so the most reliable ones will have a dominant role [108]. It is, therefore, a very efficient method to cope with comparators having heterogeneous performance. On the contrary, in conventional score-level fusion approaches (like the average of scores), each comparator is given the same weight regardless of its accuracy, a common drawback that makes the worst comparators to produce misleading results more frequently [24]. Another relevant observation is that cross-sensor error rates of the individual comparators are higher with the database captured in the same spectrum (VSSIRIS) than the database which contains images in different spectra (Cross-Eyed). As a result, there is a need to fuse more comparators with VSSIRIS to achieve maximum performance. This is an interesting phenomenon since one would expect that the comparison of images captured with visible cameras would produce better results than the comparison of near-infrared and visible images. Some authors point out that the discrepancy in colours between sensors in the visible range can be very important, leading to a significant decrease in performance when images from these sensors are compared without applying appropriate device-dependent colour corrections [45]. Since NIR images do not contain colour information, this effect may not appear in NIR–VIS comparisons.

In the present work, we use the eye corners or the sclera boundary as references to extract the periocular region of interest (ROI). While we have employed ground-truth information, an operational system would demand to locate these parts, so inaccuracies in their location would affect subsequent processing steps. In order to mitigate the effects of incorrect detection on the periocular matching performance of the different comparators and obtain a measure of their capabilities in ideal conditions [12], we have not implemented any detector of the necessary references. Even if errors in the detection will influence the overall performance of the recognition chain, feature extraction methods are not necessarily affected in the same way. This is seen for example in [109] with the iris modality, which will serve as inspiration for a similar systematic study with periocular images. The amount of periocular area around the eye necessary to provide good accuracy is another subject of study, with studies showing differences depending on the spectrum [110]. In VSSIRIS, the available images (captured with smartphones) contain a bigger periocular portion than images from the Cross-Eyed database (Fig. 7). However, it is not sufficient to provide better *cross-sensor* accuracy. Therefore, an interesting source of future research work will be to test the resilience against a variable amount of periocular area, including occlusions [12].

Another observation is that the proposed fusion method needs to be trained separately for each domain (NIR–VIS or VIS–VIS). This is not exclusive of this method but an issue that is common to score-level

fusion methods in general. Since the scores given by different systems do not necessarily lie in the same range, they are made comparable by mapping them to a common interval using score normalization techniques [20]. Even the score distributions of a given algorithm do not necessarily lie in the same range if the operational conditions are different, such as operating in NIR–VIS or VIS–VIS domains. Just changing a sensor by a more recent one from the same manufacturer may have the same effect [39], and the shape of the distributions are not necessarily equal either. One obvious effect of the difference between score distributions in different domains is that the accuracy of the comparators is different, not only in absolute numbers but also in the relative differences among them (Table 5 vs. Table 9). For example, the best comparator in Table 5 is SIFT, and it is one order of magnitude better than the others. On the other hand, in Table 9, the EER of SIFT is only a little ahead of Resnet101 or Densenet201, and the FRR is even worse. Another observable effect of this phenomenon is that the slope of the DET curves is not the same either (Fig. 9 vs. Fig. 13). For these reasons, the normalization and the fusion algorithms will usually need different training for each context. The calibration method employed implicitly finds the weight to be given to each system, so if their absolute or relative performance changes, the weights need to change accordingly. The same can be said about the other fusion algorithms evaluated. The number of systems that are needed to achieve maximum performance will not necessarily be the same either (Fig. 10 vs. Fig. 14), nor the individual systems involved in the fusion (Table 6 vs. Table 10). These observations are also backed up by a number of previous studies with different biometrics modalities [86,91,111,112]. As a future work in this direction, we are looking at the robustness of the different comparators to cross-domain training, i.e. training the calibration in one domain and testing in the other. We speculate that some comparators may be more robust than others, so using only those for calibration would allow transferring the training for one domain to the other without needing to re-train in the target domain. The use of several databases in one domain is also another way to test the generalization of the suggested approach by cross-database training [113]. As future work, we are also exploring to exploit deep learning frameworks to learn the variability between images in different spectra or captured with different sensors. One plausible approach is the use of Generative Adversarial Networks [114] to map images as if they were captured by the same sensor. This has the advantage that images can be compared using standard feature extraction methods such as the ones employed in this paper, which have been shown to work better if images are captured using the same sensor.

In the context of smartphone recognition, where high-resolution images may be available, fusion with the iris modality is another possibility to increase recognition accuracy [91]. However, it demands segmentation, which might be an issue if the image quality is not sufficiently high [15]. This motivates pursuing the periocular modality, as in the current study. We will also validate our methodology using databases not only limited to two devices or spectra, e.g. [45,52], and also including more extreme variations in camera specifications and imaging conditions, such as low resolution, illumination or pose variability. For such low-quality imaging conditions, super-resolution techniques may also be helpful [115] and will be investigated as well.

Finally, recent interest in learning biases around face recognition [116,117] motivates future research to study learning biases in the periocular region and developing new methods to reduce undesired biases [118] in that important facial region.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Part of this work was done while F. A.-F. was a visiting researcher at the Norwegian University of Science and Technology in Gjøvik (Norway), funded by EU COST Action IC1106. Authors from HH thank the Swedish Research Council (project 2016-03497), the Swedish Knowledge Foundation (CAISR and SIDUS-AIR Program), and the Swedish Innovation Agency VINNOVA (project 2018-00472) for funding his research. Authors from UAM are funded by projects: PRIMA (MSCA-ITN-2019-860315), TRESPASS-ETN (MSCA-ITN-2019-860813), and BIBECA (RTI2018-101248-B-I00 MINECO).

References

[1] F. Alonso-Fernandez, J. Bigun, A survey on periocular biometrics research, Pattern Recognit. Lett. 82 (2016) 92–105.
[2] I. Nigam, M. Vatsa, R. Singh, Ocular biometrics: A survey of modalities and fusion approaches, Inf. Fusion 26 (2015) 1–35.
[3] H. Proenca, M. Nixon, M. Nappi, E. Ghaleb, G. Ozbulak, H. Gao, H.K. Ekenel, K. Grm, V. Struc, H. Shi, X. Zhu, S. Liao, Z. Lei, S.Z. Li, W. Gutfeter, A. Pacut, J. Brogan, W.J. Scheirer, E. Gonzalez-Sosa, R. Vera-Rodriguez, J. Fierrez, J. Ortega-Garcia, D. Riccio, L.D. Maio, Trends and controversies, IEEE Intell. Syst. 33 (3) (2018) 41–67.
[4] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, F. Alonso-Fernandez, Facial soft biometrics for recognition in the wild: Recent works, annotation and cots evaluation, IEEE Trans. Inf. Forensics Secur. 13 (8) (2018) 2001–2014.
[5] P. Tome, J. Fierrez, R. Vera-Rodriguez, D. Ramos, Identification using face regions: Application and assessment in forensic scenarios, Forensic Sci. Int. (233) (2013) 75–83.
[6] P. Tome, J. Fierrez, R. Vera-Rodriguez, J. Ortega-Garcia, Combination of face regions in forensic scenarios, J. Forensic Sci. 60 (4) (2015) 1046–1051.
[7] M. Ngan, P. Grother, K. Hanaoka, Ongoing frvt part 6a: Face recognition accuracy with face masks using pre-covid-19 algorithms, NISTIR 8311 - http://www.nist.gov/itl/iad/ig/pft.cfm.
[8] B. Klare, Rank one's next-generation periocular recognition algorithm, 2020, URL https://blog.rankone.io/2020/05/06/rank-ones-next-generation-periocular-recognition-algorithm/.
[9] Face id firms battle covid-19 as users shun fingerprinting, Biometr. Technol. Today 2020 (4) (2020) 1–2.
[10] R.R. Jillela, A. Ross, V.N. Boddeti, B.V.K.V. Kumar, X. Hu, R. Plemmons, P. Pauca, Handbook of Iris Recognition, Springer, 2013, pp. 281–308, Ch. Iris segmentation for challenging periocular images.
[11] A. Jain, K. Nandakumar, A. Ross, 50 Years of biometric research: Accomplishments, challenges, and opportunities, Pattern Recognit. Lett. 79 (2016) 80–105.
[12] U. Park, R.R. Jillela, A. Ross, A.K. Jain, Periocular biometrics in the visible spectrum, IEEE Trans. Inf. Forensics Secur. 6 (1) (2011) 96–106.
[13] A. Rattani, R. Derakhshani, Ocular biometrics in the visible spectrum: A survey, Image Vis. Comput. 59 (2017) 1–16.
[14] H. Proença, J.C. Neves, Deep-prwis: Periocular recognition without the iris and sclera using deep learning frameworks, IEEE Trans. Inf. Forensics Secur. 13 (4) (2018) 888–896.
[15] F. Alonso-Fernandez, J. Fierrez, J. Ortega-Garcia, Quality measures in biometric systems, IEEE Secur. Priv. 10 (6) (2012) 52–62.
[16] L. Xiao, Z. Sun, T. Tan, Fusion of iris and periocular biometrics for cross-sensor identification, in: Proc. 7th Chinese Conference on Biometric Recognition, CCBR, 2012, pp. 202–209.
[17] M. Moreno-Moreno, J. Fierrez, J. Ortega-Garcia, Biometrics beyond the visible spectrum: Imaging technologies and applications, in: J. Fierrez, J. Ortega-Garcia, A. Esposito, A. Drygajlo, M. Faundez-Zanuy (Eds.), Proceedings of BioID-Multicomm, in: LNCS, vol. 5707, Springer, 2009, pp. 154–161.
[18] R.R. Jillela, A. Ross, Matching face against iris images using periocular information, in: Proc Intl Conf Image Processing, ICIP, 2014, pp. 4997–5001.
[19] P. Tome, R. Vera-Rodriguez, J. Fierrez, J. Ortega-Garcia, Facial soft biometric features for forensic face recognition, Forensic Sci. Int. 257 (2015) 171–284.
[20] J. Fierrez, A. Morales, R. Vera-Rodriguez, D. Camacho, Multiple classifiers in biometrics. part 1: Fundamentals and review, Inf. Fusion 44 (2018) 57–64.
[21] A. Lumini, L. Nanni, Overview of the combination of biometric matchers, Inf. Fusion 33 (2017) 71–85.
[22] M. Singh, R. Singh, A. Ross, A comprehensive overview of biometric fusion, Inf. Fusion 52 (2019) 187–205.
[23] F. Alonso-Fernandez, J. Fierrez, D. Ramos, J. Gonzalez-Rodriguez, Quality-based conditional processing in multi-biometrics: Application to sensor interoperability, IEEE Trans. Syst. Man Cybern. A 40 (6) (2010) 1168–1179.
[24] A. Jain, K. Nandakumar, A. Ross, Score normalization in multimodal biometric systems, Pattern Recognit. 38 (12) (2005) 2270–2285.

[25] B. Gutschoven, P. Verlinde, Multi-modal identity verification using support vector machines (svm), in: Proceedings of the Third International Conference on Information Fusion, Vol. 2, 2000, pp. THB3/3–THB3/8.

[26] Y. Ma, B. Cukic, H. Singh, A classification approach to multi-biometric score fusion, in: Audio- and Video-Based Biometric Person Authentication, AVBPA, 2005, pp. 484–493.

[27] A.F. Sequeira, L. Chen, J. Ferryman, F. Alonso-Fernandez, J. Bigun, K.B. Raja, R. Raghavendra, C. Busch, P. Wild, Cross-eyed - cross-spectral iris/periocular recognition database and competition, in: Proc Intl Conf of the Biometrics Special Interest Group, BIOSIG, 2016, pp. 1–5.

[28] K.B. Raja, R. Raghavendra, V.K. Vemuri, C. Busch, Smartphone based visible iris recognition using deep sparse filtering, Pattern Recognit. Lett. 57 (2015) 33–42.

[29] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proc IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1, 2005, pp. 886–893.

[30] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.

[31] D. Lowe, Distinctive image features from scale-invariant key points, Int. J. Comput. Vis. 60 (2) (2004) 91–110.

[32] F. Alonso-Fernandez, A. Mikaelyan, J. Bigun, Compact multi-scale periocular recognition using SAFE features, in: 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 1455–1460.

[33] F. Alonso-Fernandez, J. Bigun, Near-infrared and visible-light periocular recognition with gabor features using frequency-adaptive automatic eye detection, IET Biometr. 4 (2) (2015) 74–89.

[34] K.B. Raja, R. Raghavendra, C. Busch, Scale-level score fusion of steered pyramid features for cross-spectral periocular verification, in: 2017 20th International Conference on Information Fusion (Fusion), 2017, pp. 1–7.

[35] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: M.W.J. Xianghua Xie, G.K.L. Tam (Eds.), Proceedings of the British Machine Vision Conference (BMVC), BMVA Press, 2015, pp. 41.1–41.12.

[36] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778.

[37] G. Huang, Z. Liu, L.v.d. Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261–2269.

[38] K. Zuiderveld, Graphics Gems IV, 1994, pp. 474–485.

[39] F. Alonso-Fernandez, K.B. Raja, C. Busch, J. Bigun, Log-likelihood score level fusion for improved cross-sensor smartphone periocular recognition, in: 2017 25th European Signal Processing Conference (EUSIPCO), 2017, pp. 271–275.

[40] S. Pigeon, P. Druyts, P. Verlinde, Applying logistic regression to the fusion of the NIST'99 1-speaker submissions, Digit. Signal Process. 10 (2000) 237–248.

[41] N. Brummer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. van Leeuwen, P. Matejka, P. Scwartz, A. Strasheim, Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST Speaker Recognition Evaluation 2006, IEEE Trans. Audio Speech Signal Process. 15 (7) (2007) 2072–2084.

[42] R. Duda, P. Hart, D. Stork, Pattern Classification - 2nd Edition, 2004.

[43] N. Poh, T. Bourlai, J. Kittler, L. Allano, F. Alonso-Fernandez, O. Ambekar, J. Baker, B. Dorizzi, O. Fatukasi, J. Fierrez, H. Ganster, J. Ortega-Garcia, D. Maurer, A. Salah, T. Scheidat, C. Vielhauer, Benchmarking quality-dependent and cost-sensitive score-level multimodal biometric fusion algorithms, IEEE Trans. Inf. Forensics Secur. 4 (4) (2009) 849–866.

[44] J. Fierrez-Aguilar, Adapted Fusion Schemes for Multimodal Biometric Authentication (Ph.D. thesis), Universidad Politecnica de Madrid, 2006.

[45] G. Santos, E. Grancho, M.V. Bernardo, P.T. Fiadeiro, Fusing iris and periocular information for cross-sensor recognition, Pattern Recognit. Lett. 57 (2015) 52–59.

[46] K.B. Raja, R. Raghavendra, C. Busch, Dynamic scale selected laplacian decomposed frequency response for cross-smartphone periocular verification in visible spectrum, in: Proc 19th International Conference on Information Fusion (FUSION), 2016, pp. 2206–2212.

[47] C. Kandaswamy, J.C. Monteiro, L.M. Silva, J.S. Cardoso, Multi-source deep transfer learning for cross-sensor biometrics, Neural Comput. Appl. 28 (9) (2017) 2461–2475.

[48] A. Sharma, S. Verma, M. Vatsa, R. Singh, On cross spectral periocular recognition, in: Proc IEEE International Conference on Image Processing (ICIP), 2014, pp. 5007–5011.

[49] Z. Cao, N.A. Schmid, Fusion of operators for heterogeneous periocular recognition at varying ranges, Pattern Recognit. Lett. 82 (Part 2) (2016) 170–180.

[50] N.P. Ramaiah, A. Kumar, On matching cross-spectral periocular images for accurate biometrics identification, in: Proc IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), 2016, pp. 1–6.

[51] S.S. Behera, M. Gour, V. Kanhangad, N. Puhan, Periocular recognition in cross-spectral scenario, in: Proc IEEE International Joint Conference on Biometrics, IJCB, 2017, pp. 681–687.

[52] N. Vetrekar, K.B. Raja, R. Ramachandra, R. Gad, C. Busch, Multi-spectral imaging for robust ocular biometrics, in: 2018 International Conference on Biometrics (ICB), 2018, pp. 195–201.

[53] K. Hernandez-Diaz, F. Alonso-Fernandez, J. Bigun, Cross spectral periocular matching using resnet features, in: Proc International Conference on Biometrics (ICB), 2019.

[54] M.D. Marsico, C. Galdi, M. Nappi, D. Riccio, FIRME: Face and iris recognition for mobile engagement, Image Vis. Comput. 32 (12) (2014) 1161–1172.

[55] Z. Guo, L. Zhang, D. Zhang, A completed modeling of local binary pattern operator for texture classification, IEEE Trans. Image Process. 19 (6) (2010) 1657–1663.

[56] A. Oliva, A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, Int. J. Comput. Vis. 42 (3) (2001) 145–175.

[57] A. Bosch, A. Zisserman, X. Munoz, Representing shape with a spatial pyramid kernel, in: Proc. of the 6th ACM International Conference on Image and Video Retrieval, CIVR, 2007, pp. 401–408.

[58] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, W. Gao, Wld: A robust local image descriptor, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1705–1720.

[59] J. Bigun, Vision with Direction, Springer, 2006.

[60] C. Padole, H. Proenca, Periocular recognition: Analysis of performance degradation factors, in: Proc Intl Conf Biometrics, ICB, 2012, pp. 439–445.

[61] F. Smeraldi, J. Bigün, Retinal vision applied to facial features detection and face authentication, Pattern Recognit. Lett. 23 (4) (2002) 463–475.

[62] M. Bulacu, L. Schomaker, Text-independent writer identification and verification using textural and allographic features, IEEE TPAMI 29 (4) (2007) 701–717.

[63] M. Unser, N. Chenouard, D.V.D. Ville, Steerable pyramids and tight wavelet frames in l2(bbrd), IEEE Trans. Image Process. 20 (2011) 2705–2721.

[64] M.N. Do, M. Vetterli, Rotation invariant texture characterization and retrieval using steerable wavelet-domain hidden markov models, IEEE Trans. Multimed. 4 (4) (2002) 517–527.

[65] G. Tzagkarakis, B. Beferull-Lozano, P. Tsakalides, Rotation-invariant texture retrieval with gaussianized steerable pyramids, IEEE Trans. Image Process. 15 (9) (2006) 2702–2718.

[66] J. Portilla, E.P. Simoncelli, A parametric texture model based on joint statistics of complex wavelet coefficients, Int. J. Comput. Vis. 40 (1) (2000) 49–70.

[67] S. Lyu, E.P. Simoncelli, Modeling multiscale subbands of photographic images with fields of gaussian scale mixtures, IEEE Trans. Pattern Anal. Mach. Intell. 31 (4) (2009) 693–706.

[68] S.-T. Li, Y. Li, Y.-N. Wang, Comparison and fusion of multiresolution features for texture classification, in: Proc Intl Conf on Machine Learning and Cybernetics, Vol. 6, IEEE, 2004, pp. 3684–3688.

[69] M. El Aroussi, M. El Hassouni, S. Ghouzali, M. Rziza, D. Aboutajdine, Novel face recognition approach based on steerable pyramid feature extraction, in: Proc IEEE Intl Conf on Image Processing (ICIP), IEEE, 2009, pp. 4165–4168.

[70] C. Su, Y. Zhuang, L. Huang, F. Wu, Steerable pyramid-based face hallucination, Pattern Recognit. 38 (6) (2005) 813–824.

[71] E.P. Simoncelli, W.T. Freeman, The steerable pyramid: A flexible architecture for multi-scale derivative computation, in: Proc Intl Conf on Image Processing, Vol. 3, IEEE, 1995, pp. 444–447.

[72] W.T. Freeman, E.H. Adelson, et al., The design and use of steerable filters, IEEE Trans. Pattern Anal. Mach. Intell. 13 (9) (1991) 891–906.

[73] H. Greenspan, S. Belongie, R. Goodman, P. Perona, Rotation invariant texture recognition using a steerable pyramid, in: Proc IAPR Intl Conf on Pattern Recognition, Vol. 2, IEEE, 1994, pp. 162–167.

[74] V. Ojansivu, J. Heikkilä, Blur insensitive texture classification using local phase quantization, in: Image and Signal Processing, Springer, 2008, pp. 236–243.

[75] F. Alonso-Fernandez, P. Tome-Gonzalez, V. Ruiz-Albacete, J. Ortega-Garcia, Iris recognition based on sift features, in: Proc First IEEE International Conference on Biometrics, Identity and Security (BIdS), 2009, pp. 1–8.

[76] K. Nguyen, C. Fookes, A. Ross, S. Sridharan, Iris recognition with off-the-shelf cnn features: A deep learning perspective, IEEE Access 6 (2018) 18848–18855.

[77] K. Hernandez-Diaz, F. Alonso-Fernandez, J. Bigun, Periocular recognition using CNN features off-the-shelf, in: Proc Intl Conf Biometrics Special Interest Group, BIOSIG, 2018, pp. 1–5.

[78] A.S. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, Cnn features off-the-shelf: An astounding baseline for recognition, in: Proc IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2014, pp. 512–519.

[79] G.B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, Tech. Rep. 07-49, University of Massachusetts, Amherst, 2007.

[80] L. Wolf, T. Hassner, I. Maoz, Face recognition in unconstrained videos with matched background similarity, in: Proc. Conf on Computer Vision and Pattern Recognition, CVPR, 2011, pp. 529–534.

[81] N. Brummer, J. du Preez, Application independent evaluation of speaker detection, Comput. Speech Lang. 20 (2006) 230–275.

[82] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. Toledano, J. Ortega-Garcia, Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition, IEEE Trans. Audio Speech Lang. Process. 15 (7) (2007) 2104–2115.

[83] L. Ferrer, M. Graciarena, A. Zymnis, E. Shriberg, System combination using auxiliary information for speaker verification, in: Proc IEEE Intl Conf on Acoustics, Speech and Signal Processing, 2008, pp. 4853–4856.
[84] E. Bigun, J. Bigun, B. Duc, S. Fischer, Expert conciliation for multi modal person authentication systems by Bayesian statistics, in: Proc Intl Conf Audio- and Video-Based Biometric Person Authentication, in: AVBPA Springer LNCS, vol. 1206, 1997, pp. 291–300.
[85] J. Kittler, M. Hatef, R. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 20 (3) (1998) 226–239.
[86] J. Fierrez-Aguilar, L. Nanni, J. Ortega-Garcia, R. Capelli, D. Maltoni, Combining multiple matchers for fingerprint verification: A case study in FVC2004, in: Proc Int Conf Image Analysis and Processing, in: ICIAP Springer LNCS, vol. 3617, 2005, pp. 1035–1042.
[87] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag New York, Inc., New York, NY, USA, 1995.
[88] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[89] A.F. Sequeira, L. Chen, J. Ferryman, P. Wild, F. Alonso-Fernandez, J. Bigun, K.B. Raja, R. Raghavendra, C. Busch, T. de Freitas Pereira, S. Marcel, S.S. Behera, M. Gour, V. Kanhangad, Cross-eyed 2017: Cross-spectral iris/periocular recognition competition, in: IEEE International Joint Conference on Biometrics (IJCB), 2017, pp. 725–732.
[90] C. Rathgeb, A. Uhl, Secure iris recognition based on local intensity variations, in: Proc ICIAR 6112, 2010, pp. 266–275.
[91] F. Alonso-Fernandez, A. Mikaelyan, J. Bigun, Comparison and fusion of multiple iris and periocular matchers using near-infrared and visible images, in: Proc 3rd Intl Workshop on Biometrics and Forensics (IWBF), 2015, pp. 1–6.
[92] K. Hollingsworth, S.S. Darnell, P.E. Miller, D.L. Woodard, K.W. Bowyer, P.J. Flynn, Human and machine performance on periocular biometrics under near-infrared light and visible light, IEEE Trans. Inf. Forensics Secur. 7 (2) (2012) 588–601.
[93] D.L. Woodard, S.J. Pundlik, J.R. Lyle, P.E. Miller, Periocular region appearance cues for biometric identification, in: Proc IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2010, pp. 162–169.
[94] J. Fierrez-Aguilar, Y. Chen, J. Ortega-Garcia, A. Jain, Incorporating image quality in multi-algorithm fingerprint verification, in: Proc Intl Conf Biometrics, in: ICB Springer LNCS, vol. 3832, 2006, pp. 213–220.
[95] International Organization for Standardization, ISO/IEC 19795-1:2006 Biometric Performance Testing and Reporting – Part 1: Principles and Framework, JTC1/SC37/Biometrics, 2006, https://www.iso.org/standard/41447.html.
[96] A. Jain, R. Duin, J. Mao, Statistical pattern recognition: A review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (1) (2000) 4–37.
[97] P.E. Miller, J.R. Lyle, S.J. Pundlik, D.L. Woodard, Performance evaluation of local appearance based periocular recognition, in: Proc Fourth IEEE Intl Conf on Biometrics: Theory, Applications and Systems (BTAS), 2010, pp. 1–6.
[98] F. Juefei-Xu, M. Savvides, Subspace-based discrete transform encoded local binary patterns representations for robust periocular matching on NIST face recognition grand challenge, IEEE Trans. Image Process. 23 (8) (2014) 3490–3505.
[99] D.L. Woodard, S. Pundlik, P. Miller, R. Jillela, A. Ross, On the fusion of periocular and iris biometrics in non-ideal imagery, in: Proc IAPR Intl Conf on Pattern Recognition, 2010, pp. 201–204.
[100] V. Boddeti, J. Smereka, B. Kumar, A comparative evaluation of iris and ocular recognition methods on challenging ocular images, in: Proc Intl Joint Conf Biometrics, IJCB, 2011, pp. 1–8.
[101] A. Ross, R. Jillela, J. Smereka, V. Boddeti, B. Kumar, R. Barnard, X. Hu, P. Pauca, R. Plemmons, Matching highly non-ideal ocular images: An information fusion approach, in: Proc Intl Conf Biometrics, ICB, 2012, pp. 446–453.
[102] C.-W. Tan, A. Kumar, Human identification from at-a-distance images by simultaneously exploiting iris and periocular features, in: Proc Intl Conf Pattern Recognition, ICPR, 2012, pp. 553–556.
[103] P. Tome, J. Fierrez, R. Vera-Rodriguez, M. Nixon, Soft biometrics and their application in person recognition at a distance, IEEE Trans. Inf. Forensics Secur. 9 (3) (2014) 464–475.
[104] S.J. Raudys, A.K. Jain, Small sample size effects in statistical pattern recognition: recommendations for practitioners, IEEE TPAMI 13 (3) (1991) 252–264.
[105] F. Roli, J. Kittler, G. Fumera, D. Muntoni, An experimental comparison of classifier fusion rules for multimodal personal identity verification systems, in: Multiple Classifier Systems, MCS, 2002, pp. 325–335.
[106] F. Alonso-Fernandez, J. Fierrez-Aguilar, H. Fronthaler, K. Kollreider, J. Ortega-Garcia, J. Gonzalez-Rodriguez, J. Bigun, Combining multiple matchers for fingerprint verification: A case study in biosecure network of excellence, Ann. Telecommun. 62 (1–2) (2007) 62–82.
[107] S. Garcia-Salicetti, J. Fierrez-Aguilar, F. Alonso-Fernandez, C. Vielhauer, R. Guest, L. Allano, T. Doan Trung, T. Scheidat, B. Ly Van, J. Dittmann, B. Dorizzi, J. Ortega-Garcia, J. Gonzalez-Rodriguez, M.B. di Castiglione, M. Fairhurst, Biosecure reference systems for on-line signature verification: A study of complementarity, Ann. Telecommun. 62 (1–2) (2007) 36–61.
[108] J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, J. Bigun, Discriminative multimodal biometric authentication based on quality measures, Pattern Recognit. 38 (5) (2005) 777–779.
[109] H. Hofbauer, F. Alonso-Fernandez, J. Bigun, A. Uhl, Experimental analysis regarding the influence of iris segmentation on the recognition rate, IET Biometr. 5 (3) (2016) 200–211.
[110] F. Alonso-Fernandez, J. Bigun, Best regions for periocular recognition with NIR and visible images, in: Proc Intl Conf Image Processing, ICIP.
[111] P. Tome, J. Fierrez, F. Alonso-Fernandez, J. Ortega-Garcia, Scenario-based score fusion for face recognition at a distance, in: Proc IEEE Workshop on Biometrics, in Association with CVPR, 2010, pp. 67–73.
[112] F. Alonso-Fernandez, J. Fierrez-Aguilar, A. Gilperez, J. Galbally, J. Ortega-Garcia, Robustness of signature verification systems to imitators with increasing skills, in: Proc IAPR Intl Conf Document Analysis and Recognition, ICDAR.
[113] E. López-López, X.M. Pardo, C.V. Regueiro, R. Iglesias, F.E. Casado, Dataset bias exposed in face verification, IET Biometr. 8 (4) (2019) 249–258.
[114] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 27, Curran Associates, Inc., 2014, pp. 2672–2680.
[115] F. Alonso-Fernandez, R.A. Farrugia, J. Bigun, J. Fierrez, E. Gonzalez-Sosa, A survey of super-resolution in iris biometrics with evaluation of dictionary-learning, IEEE Access 7 (2019) 6519–6544.
[116] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, C. Busch, Demographic bias in biometrics: A survey on an emerging challenge, IEEE Trans. Technol. Soc. 1 (2) (2020) 89–103.
[117] P. Terhörst, J.N. Kolf, M. Huber, F. Kirchbuchner, N. Damer, A. Morales, J. Fierrez, A. Kuijper, A comprehensive study on face recognition biases beyond demographics, 2021, arXiv:2103.01592.
[118] A. Morales, J. Fierrez, R. Vera-Rodriguez, R. Tolosana, Sensitivenets: Learning agnostic representations with application to face images, IEEE Trans. Pattern Anal. Mach. Intell. 43 (6) (2021) 2158–2164.