scientific reports

Check for updates

OPEN

A new periocular dataset collected by mobile devices in unconstrained scenarios

Luiz A. Zanlorensi[1]✉, Rayson Laroca[1], Diego R. Lucio[1], Lucas R. Santos[1], Alceu S. Britto Jr.[2] & David Menotti[1]

Recently, ocular biometrics in unconstrained environments using images obtained at visible wavelength have gained the researchers' attention, especially with images captured by mobile devices. Periocular recognition has been demonstrated to be an alternative when the iris trait is not available due to occlusions or low image resolution. However, the periocular trait does not have the high uniqueness presented in the iris trait. Thus, the use of datasets containing many subjects is essential to assess biometric systems' capacity to extract discriminating information from the periocular region. Also, to address the within-class variability caused by lighting and attributes in the periocular region, it is of paramount importance to use datasets with images of the same subject captured in distinct sessions. As the datasets available in the literature do not present all these factors, in this work, we present a new periocular dataset containing samples from 1122 subjects, acquired in 3 sessions by 196 different mobile devices. The images were captured under unconstrained environments with just a single instruction to the participants: to place their eyes on a region of interest. We also performed an extensive benchmark with several Convolutional Neural Network (CNN) architectures and models that have been employed in state-of-the-art approaches based on Multi-class Classification, Multi-task Learning, Pairwise Filters Network, and Siamese Network. The results achieved in the closed- and open-world protocol, considering the identification and verification tasks, show that this area still needs research and development.

Biometric systems that use ocular images have been extensively investigated due to the high level of singularity in the iris and because the periocular region can provide discriminative patterns even in noisy images[1–6]. The term ocular comprises the periocular and iris regions[7]. The periocular region comprises eyebrows, eyelashes and eyelids, while the iris is the colored region between the sclera and pupil. There are two main modes that an ocular biometric system can operate: identification (1:*N* comparison) and verification (1:1 comparison). The identification task consists of determining a subject's identity, whereas the verification one verifies whether a subject is who she/he claims to be. There are also two main protocols to evaluate biometric systems: closed-world and open-world[8,9]. In the former, the training and test sets have different samples from exactly the same subjects. On the other hand, in the open-world protocol, the training and test sets must have samples from different subjects. With these modes and protocols, it is possible to evaluate some characteristic of biometric approaches to produce discriminative features and generalization capability.

Nowadays, with the advancement of deep learning-based techniques, several methodologies applying them to ocular images have been proposed for several tasks, for example, spoofing detection[24,25], iris and periocular region detection[26–28], iris and sclera segmentation[29,30], and iris and periocular recognition[31–37]. The advancement of these technologies can be observed by the recent contests that have been conducted to evaluate the evolution of the state-of-the-art methods for different applications, such as iris recognition in heterogeneous lighting conditions (NICE.I and NICE.II)[21,38], iris recognition using mobile images (MICHE.I and MICHE.II)[2,16], iris and periocular recognition in cross-spectral scenarios (Cross-Eyed 1 and 2)[17,18], and periocular recognition using mobile images captured in different lighting conditions (VISOB 1 and 2)[23]. Note that all these contests used datasets containing images obtained in the visible wavelength. The most recent contests also used images captured by mobile devices[2,23]. The results achieved by the proposed methods have shown that it is challenging to develop a robust biometric system in such conditions, mainly due to the high intra-class variability. Based on recent works[2,5,7],

[1]Department of Informatics, Federal University of Paraná (UFPR), Curitiba 81530-000, Brazil. [2]Postgraduate Program in Informatics, Pontifical Catholic University of Paraná (PUCPR), Curitiba 80215-182, Brazil. ✉email: lazjunior@inf.ufpr.br

Dataset	Subjects	Images	Sessions	Capture devices
VSSIRIS[10]	28	560	1	2
CSIP[11]	50	2004	N/A	7
QUT[12]	53	212	N/A	2
IIITD[13]	62	1240	N/A	3
UPOL[14]	64	384	N/A	1
UTIRIS[15]	79	1540	2	2
MICHE-I[16]	92	3732	2	3
CROSS-EYED[17,18]	120	3, 840	N/A	2
PolyU Cross-Spectral[19]	209	12, 540	2	2
UBIRIS.v1[20]	241	1877	2	1
UBIRIS.v2[21]	261	11, 102	2	1
UBIPr[22]	261	10, 950	2	1
VISOB[23]	550	**158,136**	2	3
UFPR-Periocular	**1122**	33, 660	**3**	**196**

Table 1. Comparison of the available periocular datasets containing visible (VIS) images with our dataset (UFPR-Periocular). Significant values are given in bold.

we can state that developing an ocular biometric system that operates in unconstrained environments is still a challenging task, especially with images obtained by mobile devices. In this condition, the images captured by the volunteer may present several variations caused by occlusion, pose, eye gaze, off-angle, distance, resolution, and image quality (affected by the mobile device).

With the existing periocular datasets, it is difficult to assess the scalability performance of biometric applications, i.e., if an approach can produce discriminative features even in a large dataset in terms of the number of subjects. As we can see in Table 1, the datasets in the literature do not present a large number of subjects and have few capture devices and session captures. As described in some previous works[5,6], one common problem in ocular biometric systems is the within-class variability, which is generally affected by noises and attributes present in the same individual images. A robust biometric system must handle images obtained from different capture devices, extracting distinctive representations regardless of the source and environments. In this sense, samples from the same subject obtained in different sessions are of paramount importance to capture the intra-class variation caused by various noise factors.

Considering the above discussion, in this work, we introduce a new periocular dataset, called *UFPR-Periocular*. The subjects themselves collected the images that compose our dataset through a mobile application (app). In this way, the images were captured in unconstrained environments, with a minimum of cooperation from the participant, and have real noises caused by poor lighting, occlusion, specular reflection, blur, and motion blur. Figure 1 shows some samples from the UFPR-Periocular. As part of this work, we also present an extensive benchmark, employing several state-of-the-art architectures of CNN models that have been explored to develop ocular (periocular and iris) recognition biometric systems. Face and eye detection are not covered in this work. The recognition methods are evaluated with manually pre-processed images (also available in the dataset).

Note that our dataset is the largest one in terms of the number of subjects, sessions, and capture devices, as shown in Table 1. It also has more images than all datasets except VISOB. Another key feature is that the proposed dataset has images captured by 196 different mobile devices. The samples captured with less cooperation of the participant in unconstrained environments have several variations on the ocular images since they are obtained during three different sessions. To the best of our knowledge, this is the first periocular dataset with more than 1, 000 subject samples and the largest one in different capture devices in the literature. Thus, we believe that it can provide a new benchmark to evaluate and develop new robust periocular biometric approaches.

Recently, with the advancement of devices enabling the self-capture of images that can be used as biometrics, the term "selfie biometrics" has been extensively explored by the research community[39,40], especially in face and iris recognition[41–43]. As described by Rattani et al. [3], the term "selfie biometrics" consists of a biometric system where the input data is acquired by the user using the capture devices available in their device. Thus, we can consider the UFPR-Periocular dataset, presented in this work, as a selfie biometric dataset since its images were acquired by the users through their own smartphones.

The remainder of this work is organized as follows. In "Related work", we describe the periocular datasets containing VIS images for periocular biometrics. In "Dataset", we present information about the UFPR-Periocular dataset and the proposed protocol to evaluate biometric systems. "Benchmark" presents the CNN architectures used to perform the benchmark. In "Results and discussion", we present and discuss the benchmark results. Finally, the conclusions are given in "Conclusion".

Related work

In recent years, several ocular contests and datasets have been released to evaluate state-of-the-art methods for many applications. Zanlorensi et al.[7] detailed and described several datasets and contests for iris and periocular recognition. Different problems have been addressed by the researchers, such as ocular recognition in

Figure 1. Sample images from the UFPR-Periocular dataset. Observe that there is great diversity in terms of lighting conditions, age, gender, eyeglasses, specular reflection, occlusion, resolution, eye gaze, and ethnic diversity.

unconstrained environments, ocular recognition on cross-spectral scenarios, iris/periocular region detection, iris/periocular region segmentation, and sclera segmentation[44].

Existing periocular datasets can be organized into constrained (or controlled) or unconstrained (or non-controlled) environments. The quality of the images is different in constrained and unconstrained environments, as some noise can occur in the images captured in unconstrained environments such as lighting variation, occlusion, blur, specular reflection, and distance. Images can also be acquired cooperatively and non-cooperatively in relation to some image capture restrictions imposed on the subject. Ocular non-cooperative images can have some problems caused by off-angle, focus, distance, motion blur, and occlusions by some attributes such as eyeglasses, contact lenses, and makeup.

As described in[7], datasets containing images obtained at the Near-infraRed (NIR) wavelength were created mainly to investigate the intricate patterns present in the iris region[45,46]. There are also other studies on NIR ocular images, such as generating synthetic iris images[47,48], spoofing and liveness detection[49–52], contact lens detection[53–56], and template aging[57,58]. The use of NIR ocular images captured in controlled environments by biometric systems has been studied for several years. Thus, it can be considered a mature technology that has been successfully employed in several applications[3,45,46,59,60].

In general, better results can be achieved on biometric methods using VIS images by exploring the periocular region instead of the iris trait, as the iris is rich in melanin pigment that absorbs the most visible lights—not reflecting the iris features as occur with NIR lights[59]. Also, the small resolution of ocular images is a common problem that makes it almost impracticable to use the iris trait alone. Regarding these problems, the use of VIS ocular images captured in a non-cooperative way under unconstrained environments became a recent challenge. In this sense, several studies have been carried out on periocular biometric recognition using images obtained by mobile devices in uncontrolled environments using different capture devices[10,16,23]. The following datasets were developed to investigate the use of iris and periocular traits in VIS images: UPOL [14], UBIRIS.v1 [20], UBIRIS.v2 [21] and UBIPr [22]. There are also datasets of iris and periocular region images for cross-spectral recognition, i.e., match ocular images obtained at different wavelengths (NIR against VIS and vice-versa): UTIRIS [15], IIITD Multi-spectral Periocular [13], PolyU Cross-Spectral [19], CROSS-EYED [17,18], and QUT Multispectral Periocular [12]. Focusing specifically on ocular recognition using non-cooperative images obtained in uncontrolled environments by mobile devices, we highlight the following datasets: MICHE-I [16], VSSIRIS [10], CSIP [11] and VISOB [23].

Nowadays, it is difficult to evaluate the scalability factor of the state-of-the-art biometric approaches due to the size in terms of subjects and images on the available datasets. As shown in Table 1, the most extensive dataset regarding subjects and images is VISOB [23], which has 158, 136 images from 550 subjects. The ICIP 2016 Competition on mobile ocular biometric recognition[23] employed this dataset, and in the WCCI/IJCNN2020 challenge (VISOB 2.0 Dataset and Competition results available at https://sce.umkc.edu/research-sites/cibit/dataset.html), a second version of the dataset was launched. Both contests evaluated the periocular recognition using VIS images obtained by mobile devices. The second contest's main difference is that the input images were a stack with five periocular images belonging to the same subject. The best methods achieved an EER of 0.06% and 5.26% on the first and second contests, respectively.

Also using VIS ocular images, other contests were carried out to evaluate iris and periocular recognition: NICE.II[38], MICHE.II[2], and CROSS-EYED I[17] and II[18]. The NICE.II contest evaluated iris recognition using

images containing noise within the iris region. The winner method fused features extracted from the iris and the periocular region using ordinal measures, color histograms, texton histograms, and semantic information. The MICHE.II contest also evaluated iris and periocular recognition, but using images captured by mobile devices. The winner approach extracted features from the iris and the periocular region, using the rubber sheet model normalization[61] and 1-D Log-Gabor filter and Multi-Block Transitional Local Binary Patterns, respectively. Lastly, the CROSS-EYED I and II contests evaluated iris and periocular recognition on the cross-spectral scenario. In both contests, the winner approach employed handcrafted features based on Symmetry Patterns (SAFE), Gabor Spectral Decomposition (GABOR), Scale-Invariant Feature Transform (SIFT), Local Binary Patterns (LBP), and Histogram of Oriented Gradients (HOG).

Inspired by impressive results achieved by deep learning-based techniques in multiple domains[62], several methods proposing and applying such techniques have been developed to address different tasks using ocular images[4–6,24–37]. Also, as found in the literature, deep learning frameworks for ocular biometric systems are a recent technology that still needs improvement[7]. The use of ocular datasets containing images captured by mobile devices in unconstrained environments is a challenging task that has gained attention in recent years[2,5,7,23,63].

Dataset

The UFPR-Periocular dataset was created to obtain images in unconstrained scenarios that contain realistic noises caused by occlusion, blur, and variations in lighting, distance, and angles. To this end, we developed a mobile application (app) enabling the participants to collect their pictures using their smartphones (Project approved by the Ethics Committee Board from the Health Science Sector of the Federal University of Paraná, Brazil—Process CAAE 02166918.2.0000.0102, registered in the *Plataforma Brazil* system—https://plataformabrasil.saude.gov.br/). We confirm that all methods were carried out following relevant guidelines and regulations by the Ethics Committee Board from the Health Science Sector of the Federal University of Paraná. Furthermore, we confirm that an informed consent form has been obtained from all subjects, and we do not store any data that could be used to identify the subject. We confirm that all periocular images presented in this paper (Figs. 1, 3, 4, 5, 6, 7, and 10) were extracted from the UFPR-Periocular dataset and that we have permission to publish these images in open access journal. The single instructions to the participants is to place their eyes on a region of interest marked by a rectangle drawn in the app, as illustrated in "Picture" in Fig. 3. We also restricted the images to be captured in 3 sessions, with 5 images per session and a minimum interval of 8 hours between sessions. In this way, we guarantee that the dataset has samples of the same subject with different noises, mainly due to different lighting and environments. Furthermore, imposing this minimum time interval between sessions, it is possible to collect different attributes in the periocular region of the same subject, as the images are captured at different times of the day, e.g., subjects wearing and not wearing glasses and makeup. Another attractive feature of this dataset is that all participants are Brazilian, and as Brazil has great ethnic diversity, there are images of subjects from different races, making this one of the first periocular datasets with such cultural diversity.

The images were collected from June 2019 to January 2020. The gender distribution of the subjects is (53.65%) male and (46.35%) female, and approximately 66% of the subjects are under 31 years old. In total, the dataset has images captured from 196 different mobile devices—the five most used device models were: *Apple iPhone 8* (4.1%), *Apple iPhone 9* (3.1%), *Xiaomi Mi 8 Lite* (3.0%), *Apple iPhone 7* (3.0%), and *Samsung Galaxy J7 Prime* (2.7%).

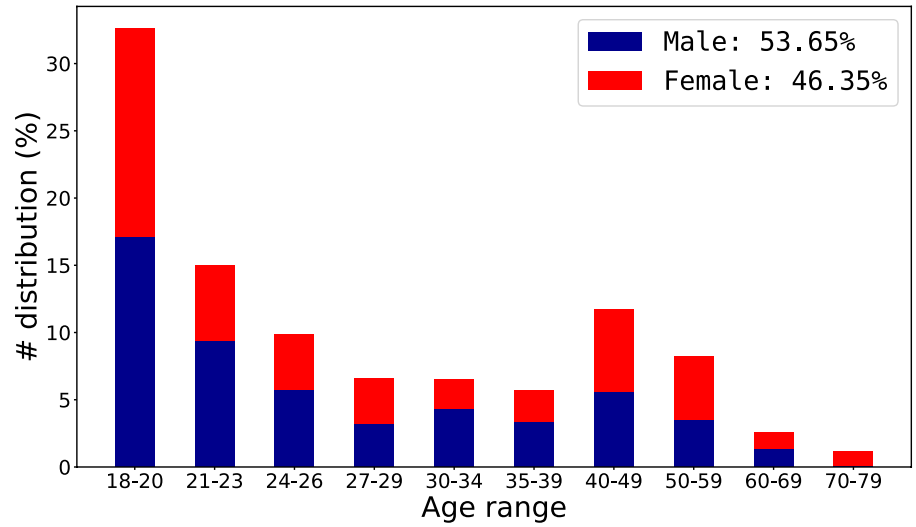
We remark that each subject captured all of their images using the same device model. The distribution of age, gender, and image resolutions present in our dataset is shown in Fig. 2.

The dataset has 16, 830 images of both eyes from 1, 122 subjects. Image resolutions vary from 360×160 to 1862×1008 pixels—depending on the mobile device used to capture the image. We cropped/separated the periocular regions of the right and left eyes to perform the benchmark, assigning a unique class to each side. Note that, once the image was cropped, the remainder image region was discarded as claimed in our project request to the Ethics Committee Board to preserve at maximum the identity of the participants. We manually annotated the eye corners with four points per image (inside and outside eye corners) and used these points to normalize the periocular region regarding scale and rotation. This process is detailed in Fig. 3.

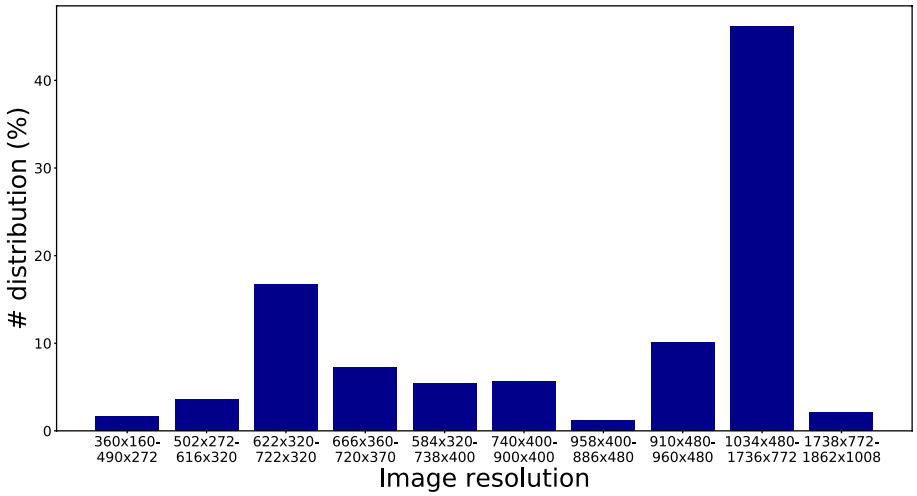
Using the center point of each eye (average corners point), the images were rotated and scaled to normalize the eye positions in a size of 512×256 pixels. Then, the images were split into two patches (256×256 pixels) to create the left and right eye sides, generating 33, 660 periocular images from 2, 244 classes. The intra- and inter-class variability in this dataset is mainly caused by lighting, occlusion, specular reflection, blur, motion blur, eyeglasses, off-angle, eye-gaze, makeup, and facial expression.

This new periocular dataset is the main contribution of this work. It can be employed in future works to evaluate and perform research in biometrics, including recognition, detection and segmentation. Furthermore, it can also be used to explore studies on recent topics such as gender and age bias[64,65], and to assess the scalability of biometric systems since this dataset is the largest one in the literature in terms of the number of subjects. Regarding the semantic segmentation problem, we reproduced the experiments presented by Banerjee et al.[66], which were proposed to generate segmentation masks for iris detection. This method consists of first transforming the raw image into the HSV and YCbCr color spaces, then using a threshold to binarize both images (HSV mask and YCbCr mask), and finally applying a dot product in both masks to generate the final global mask. However, as the images from our dataset have considerably more noise than those employed in the original work, the method could not obtain masks of satisfactory quality for us to consider them as ground truth. For this reason, the semantic segmentation problem will be addressed in future work.

Experimental protocols. We propose protocols for the two most common tasks in biometric systems: identification (1:*N*) and verification (1:1). The identification task consists of determining a subject sample iden-



(a) gender distribution among the age ranges



(b) image resolutions grouped into 10 intervals

Figure 2. Age, gender and image resolution distributions in the UFPR-Periocular dataset. (**a**) note that gender has a balanced distribution, but the age range is concentrated under 30 years old (64% of the subjects). (**b**) more than 45% of the images have a resolution between 1034 × 480 and 1736 × 772 pixels, and more than 65% of the images have resolution higher than 740 × 400 pixels.

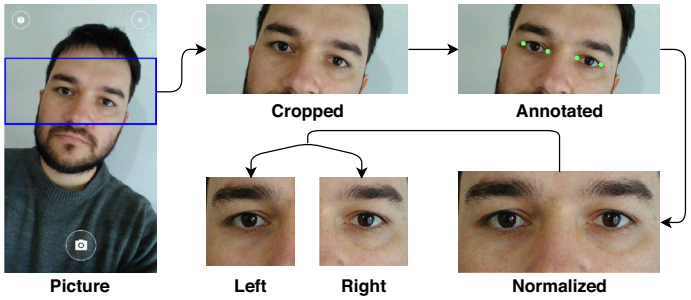


Figure 3. Image acquisition and normalization process. First, after the subject took the shot, the rectangular region (outlined in blue) was cropped and stored. Then, the images were normalized in terms of rotation and scale using the manual annotations of the corners of the eyes. Lastly, the normalized images were cropped, generating the periocular regions of the left and right eyes.

Protocol	Train/val	Images/classes			Genuine pairs/impostor pairs		
		Train	Validation	Test	Train	Validation	Test
CW	CW/CW	13, 464/2244	8976/2244	11, 220/2244	33, 660/90, 599, 256	13, 464/40, 266, 336	22, 440/12, 583, 230
OW	OW/CW	13, 464/1496	8976/1496	11, 220/748	53, 856/90, 579, 060	22, 440/40, 257, 360	78, 540/4, 190, 670
OW	OW/OW	15, 000/1000	7440/496	11, 220/748	105, 000/112, 387, 500	52, 080/27, 621, 000	78, 540/4, 190, 670

Table 2. Images, classes, and pairwise comparison distributions for the closed-world (CW) and open-world (OW) protocols. Values for each fold (3 folds).

tity (probe) within a known dataset or a cluster (gallery). The probe is compared against all the gallery samples, considering the closest match as the subject's identity. Furthermore, probabilistic models can be employed/trained using the gallery data to determine the probe subject's identity based on the highest confidence output. The verification task refers to the problem of verifying whether a subject is who she/he claims to be. If two samples match sufficiently, the identity is verified; otherwise, it is rejected[59]. Verification is usually used for positive recognition, where the goal is to prevent multiple people from using the same identity. The identification is a critical component in negative recognition, where the goal is to prevent a single person from using multiple identities[67]. Furthermore, the proposed protocol also encompasses two different scenarios: closed-world and open-world. In the closed-world protocol, the dataset is split through different samples from the same subject, i.e., training and test sets have samples of the same subjects. In the open-world protocol, there are different subjects both in the training and test sets. The identification task is performed in the closed-world protocol, while the verification task can be performed in both closed and open-world protocols. In the open-world protocol, we also propose two different splits regarding the training and validation sets. Note that we do not change the test set, keeping it in the open-world protocol, and only vary the training protocols. The first split uses the closed-world protocol, in which the training and validation sets have samples from the same subjects. The second split, on the other hand, has different subjects in the training and validation sets, i.e., in an open-world protocol. With these two training/validation splits, it is possible to use multi-class networks (classification/identification) and also models based on the similarity of two distinct inputs (verification task): Siamese networks, triplet networks, and pairwise filters. Although models built for the verification task can be trained through the closed-world protocol, the design can be better improved using the open-world protocol to split the training and validation sets, as it is a more realistic scenario regarding the test set. Table 2 summarizes the proposed protocols.

We defined 3 folds with a stratified split into training, validation, and test sets for both biometric tasks (identification and verification) for all protocols. The test set comprises all against all comparisons for genuine pairs and aiming to reduce the pairwise comparisons only impostor pairs using the images of all subjects with the same sequence index, i.e., the i-th images of each subject are combined two at-a-time to generate all impostor pairs, for $1 \leq i \leq n$, where $n = 3$ sessions $\times$ 5 images. As the UFPR-Periocular dataset has images captured under 3 sessions, we designated one session as a test set for each fold in the *closed-world protocol*. Thus, we have images from sessions 1 and 2, 2 and 3, 3 and 1 for training/validation, and sessions 3, 1, and 2 for testing, respectively for each of the three folds. To evaluate the ability of the models to recognize subjects samples at different environments, for all folds, we employed samples of both sessions in the training and validation sets to fed the models with images from the same subject varying the capture conditions. For each subject, we employed the first 3 images of each session for training and the remaining 2 for validation (60%/40% for training/validation splits). The test set contains new images from the subjects present in the training/validations sets with different noises caused by the environment, lighting, occlusion, and facial attributes.

For the *open-world protocol* we generate the training, validation, and test sets by splitting the dataset through different subjects. Thus, for each fold, the test set has samples of subjects not present in the training/validation set. Splitting sequentially by the subject index for each fold, we have samples of 748 subjects for training/validation and 374 subjects for testing. Moreover, we propose two different splits for the training/validation splits, the first one containing images of the same subject in the training and validation sets (closed-world validation). The second one contains samples from different subjects in the training and validation sets (open-world validation). Both training/validation protocols have pros and cons. The advantage of using the closed-world validation is that the training has samples of more subjects than the open-world validation protocol. However, in this scenario, the models can only learn distinctive features for the gallery samples and may not extract distinctive features for subjects not present in the training process. On the other hand, the open-world validation has samples of fewer subjects than the closed-world validation protocol, presenting a more realistic scenario since samples of subjects not known in the training stage are present in the validation set. In the closed-world validation protocol, for each one of the 748 subjects in the training set, we used the first 3 images of each session for training, and the remaining 2 for validation (60%/40% for training/validation splits). In the open-world validation protocol, we employed samples of the first 700 subjects for training and samples of the remaining 48 subjects to validate each fold. The number of the generated pairwise comparison for all protocols are detailed in Table 2. The files determining all splits and setups detailed in this section are available along with the UFPR-Periocular dataset.

Benchmark

To carry out an extensive benchmark, we employ different models and strategies based on deep learning that achieved promising results in the ImageNet dataset/contest[68] and were applied in recent works of ocular recognition[6,32,35,36,69]. These methods differ from each other in network architecture, loss function, and training

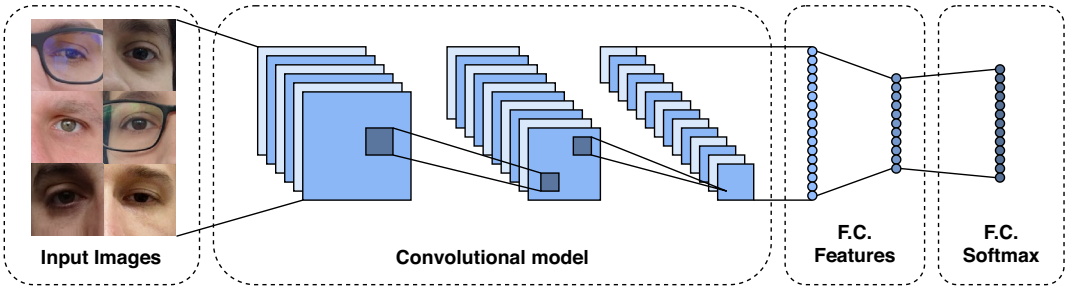


Figure 4. Multi-class classification CNN architecture.

strategies. We employed the following CNN models: Multi-class classification, Multi-task learning, Siamese networks, and Pairwise filters networks. Please note that we did not evaluate detection in this paper. We employ the images already cropped and resized (to normalize distance and rotation) to evaluate the recognition methods. In the following subsections, we describe and detail each one of them.

Multi-class classification. Multi-class classification is the task of classifying instances into three or more classes, where each sample must have a single unique class/label. Several techniques[70–72] have been proposed combining multiple binary classifiers to solve multi-class classification problems. Deep learning-based approaches usually address this problem through CNN models with softmax cross-entropy loss. Therefore, we start by evaluating several CNN architectures that achieved expressive results in the ImageNet dataset/contest[68]. In summary, the architecture of these models has several convolutional, pooling, activation, and fully-connected layers, as shown in Fig. 4.

In the training stage, a batch of images and their labels feed these models. The model extracts the image features through convolutional, pooling, and fully connected (dense) layers. The last layer is composed of a fully connected layer using the softmax cross-entropy as a loss function. In this work, following previous approaches[21,73,74], we considered each eye of each subject as a unique class, i.e., the left and right eyes belong to different classes. In this way, as expected, a person's identity can only be verified by the same eye side, i.e., the left and right eyes of the same person can not be matched. Below we describe the main characteristics of each model.

VGG. The VGG model, proposed by Simonyan and Zisserman[75], consists of a CNN using small convolution filters (3×3) with a fixed stride of 1 pixel. The spatial polling is computed by 5 max-pooling layers over a 2×2 pixel window. Two models were proposed varying the number of convolutional layers: VGG16 and VGG19. Both models have two fully connected layers at the top with 4096 channels each—these architectures achieved the first and second places in the localization and classification tracks on the ImageNet Challenge 2014. The authors also stated that it is possible to improve prior-art configurations by increasing the depth of the models. Parkhi et al.[76] applied these models (called VGG16-Face) on the face recognition problem, showing that a deep CNN with a simpler network architecture can achieve results comparable to the state of the art. Furthermore, recent approaches for ocular (iris/periocular) biometrics employing VGG models have demonstrated the ability to produce discriminant features[6,32,35,36,69,77,78]. In this work, we employed the VGG16 and VGG16-Face to perform the benchmark.

ResNet. The Residual Network (ResNet) was introduced by He et al.[79] and applied to biometrics for face recognition[80], iris recognition[6,35,69,77,81] and periocular recognition[6,37,78,82]. The authors addressed the degradation (vanishing gradient) problem caused by deeper network architectures proposing a deep residual learning framework. They added shortcut connections between residual blocks to insert residual information. These residual blocks are composed of a weighted layer followed by batch normalization, an activation function, another weighted layer, and batch normalization. Let $F(x)$ be a residual block, and x the input of this block (identity map), the residual information consists of adding x to $F(x)$, i.e., $F(x) + x$, and using it as input to the next residual block. Different architectures were proposed and evaluated, varying the depth of the models: ResNet50, ResNet101, and ResNet152. These models achieved promising results on the ImageNet dataset[68]. In[83], He et al. proposed the ResNetV2 by changing the residual block by adding a pre-activation into it. Empirical experiments showed that the proposed method improved the network generalization ability, reporting better results than ResNetV1 on ImageNet.

InceptionResNet. The InceptionResNet model[84], combines the residual connections[79] and the inception architecture[85]. The first inception model[86], known as GoogLeNet, introduced the Inception module aiming to increase the network depth while keeping a relatively low computational cost. The main idea of inception is to approximate a sparse CNN with a normal dense construction. The inception module consists of several convolutional layers, where their output filter banks are concatenated and used as the input to the next module. The model version difference is based on the organization inside its inception module. Combining the residual connections with the InceptionV3 and InceptionV4 models, the author developed InceptionResNetV1 and InceptionResNetV2, respectively. Experiments performed on the ImageNet dataset showed that the InceptionResNet

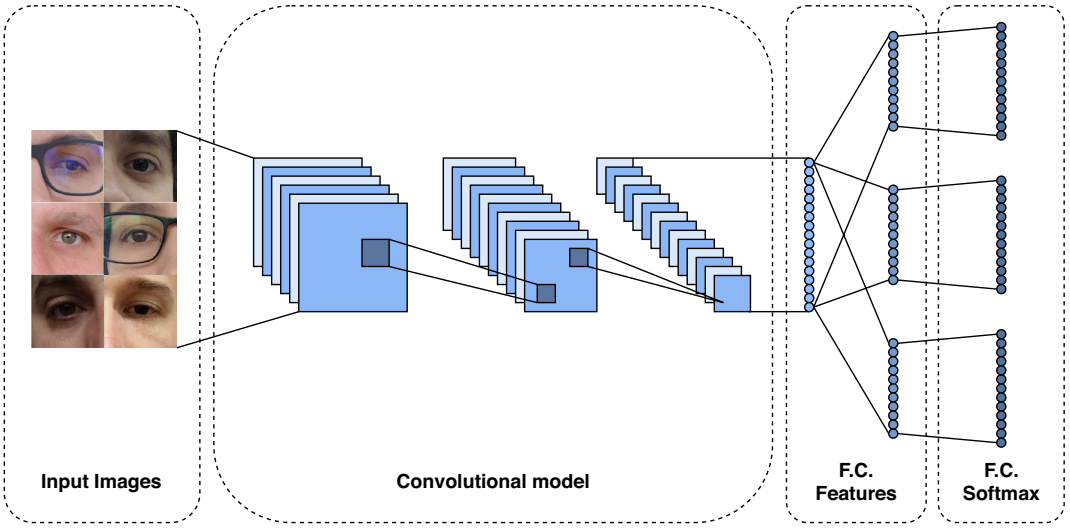


Figure 5. Multi-task CNN architecture. In this model, each task has its own output and all tasks share the convolutional layers. The loss of all tasks is used to update the weights of the convolutional layers.

models trained faster and reached slightly better results than the inception architecture[84]. In our experiments, we employed the InceptionResNetV2 model since it achieved the best results on ImageNet.

MobileNet. The first version of the MobileNet model (MobileNetV1)[87] was developed focusing on mobile and embedded vision applications, in which it is desirable that the CNN model has a small size and high computational efficiency. This model is based on depthwise separable filters, which are composed of depthwise and pointwise convolutions. As described in[87], depthwise convolutions apply a single filter for each input channel, and pointwise convolutions use a 1×1 convolution to compute a linear combination of the depthwise output. Both layers use batch normalization and ReLU activation. MobileNetV1 achieved promising results in both terms of performance and accuracy on several tasks such as fine-grained recognition, large scale geolocation, face attributes classification, object detection, and face recognition[87]. MobileNetV2[88] combines the first version architecture with an inverted ResNet[79] structure, which has shortcut connections between the bottleneck layers. Experiments performed in different tasks such as image classification, object detection, and image segmentation showed that the MobileNetV2 can achieve high accuracy with low computation costs compared to state-of-the-art methods[88].

DenseNet. The Dense Convolutional Network (DenseNet) model[89] consists of a CNN architecture where each layer is connected to every other layer in a feed-forward way. Thus, let L be the number of layers from a network, a DenseNet layer has $\frac{L(L+1)}{2}$ direct connections with subsequent layers—instead of L as a traditional CNN model. As in the ResNet models[79,83], these connections can handle the vanishing-gradient problem and ensure maximum information flow between layers. The feed-forward is preserved, passing the output from all layers as an additional input to the subsequent ones in a channel-wise concatenation. The DenseNet models achieved state-of-the-art accuracies in image classification on the CIFAR10/100 and ImageNet datasets[68,89]. The authors proposed different models varying the depth of the network. In our experiments, we employed DenseNet121 (the shallowest one).

Xception. Xception model was inspired by inception modules, being defined as an intermediate step between convolution and depthwise separable convolution operation[90]. The proposed architecture replaces the standard inception modules with depthwise separable convolutions and residual connections. The Xception is similar to InceptionV3 in terms of parameters but outperforms it on the ImageNet dataset[68].

Multi-task learning. Multi-task learning improves generalization using the domain information of related tasks as an inductive bias[91]. This architecture learns several tasks using a shared CNN model, where each task can help the generalization of other tasks. Caruana[91] introduced the Multi-task learning concept and evaluated it in different domains, demonstrating that this method can achieve better results than single-task learning models for related tasks. In deep neural networks, multi-task learning can be performed by two different setups: hard or soft parameter sharing[92]. All the hidden (convolutional) layer weights are shared in the hard parameter sharing, i.e., the model learns a single representation for all tasks. In this configuration, it is also possible to add specific layers for different tasks[93]. On the other hand, each task is processed by a different model in the soft parameter sharing. Then, the parameters of these models are regularized to encourage similarities among them.

As shown in Fig. 5, our Multi-task network shares all convolutional layers and some dense layers. The model has exclusive dense layers for each task, followed by the prediction layers, using the softmax cross-entropy as function loss.

#	Layer	Connected to	Input	Output
0	MobileNetV2 (88 layers)	–	224 × 224 × 3	1280
1	Dense (classes)	#0	1280	256
2	Dense (age)	#0	1280	256
3	Dense (gender)	#0	1280	256
4	Dense (eye side)	#0	1280	256
5	Dense (smartphone model)	#0	1280	256
6	Predict (classes)	#1	256	2244
7	Predict (age)	#2	256	10
8	Predict (gender)	#3	256	2
9	Predict (eye side)	#4	256	2
10	Predict (smartphone model)	#5	256	196

Table 3. Multi-task architecture in the closed-world protocol.

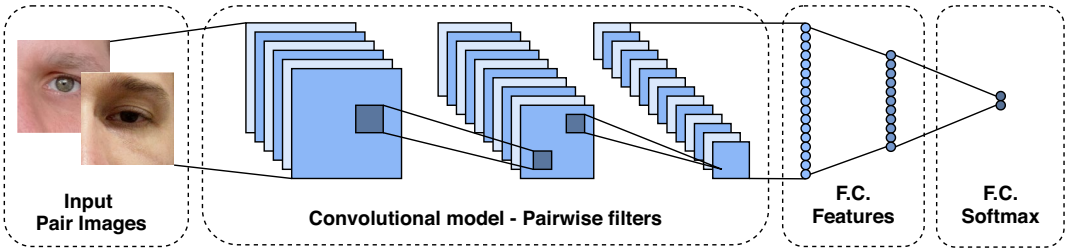


Figure 6. Pairwise filters CNN architecture. This model contains filters that directly learn the similarity between a pair of images. The output informs whether the images are of the same person or not.

In this work, based on the results of multi-class classification, we employ MobileNetV2 as the base model on our multi-task approach. Furthermore, as detailed in Table 3, we build our multi-task model with hard parameter sharing for the following 5 tasks: (i) class prediction, (ii) age rate, (iii) gender, (iv) eye side, and (v) smartphone model.

For the age estimation task, we generate the classes by grouping ages into the following 10 ranges: 18–20, 21–23, 24–26, 27–29, 30–34, 35–39, 40–49, 50–59, 60–69, and 70–79. The gender and eye side prediction tasks have only 2 classes, while the smartphone model prediction has 196 classes. Note that it is possible to employ weighted loss for each task in the Multi-task learning networks, penalizing the wrong classification of some tasks more than others. For simplicity, in this work, we do not use weighted losses in our experiments, giving equal importance to all tasks.

As shown in Table 3, we build exclusive dense layers for each task by connecting them directly to the backbone model (MobileNetV2). Then, each dense layer is connected to its respective prediction layer, making it possible that each task has its own specialized (feature) dense layer.

Pairwise filters network. Inspired by Liu et al.[94], which is one of the first works applying deep learning for iris verification, we also evaluate the performance of the pairwise filters network. This kind of model directly learns the similarity between a pair of images through pairwise filters. The Pairwise Filters Network is a Multi-class classification model that contains one or two outputs informing whether the input pairs are from the same class or from different classes. The difference is that the network input is a pair of images instead of a single image. Thus, the network architecture consists of convolutional, pooling, activation, and fully connected layers, as shown in Fig. 6.

As described by Liu et al.[94], in this kind of model the similarity map is generated through convolution and summarizes the feature maps of a pair of input images. We generate the input pairs by concatenating the images at their channel levels. Let two RGB images with shapes of 224 × 224 × 3, concatenating both images by their channels; the resulting input image will have a shape of 224 × 224 × 6 (224 × 244 pixels by 6 channels, 3 from the first image and 3 from the second image). These images proceed through convolution layers that generate feature maps regarding their similarity. The output of our model has two neurons and uses a softmax cross-entropy loss. As the verification problem has only two classes, this model's output can have only one neuron using a binary cross-entropy loss function. As in the Multi-task network, we employ MobileNetV2 as the base model for our Pairwise Filters Network.

Siamese network. Siamese networks were first described by Bromley et al.[95] for signature verification. This architecture consists of twin branches sharing their trainable parameters. Such models are generally employed for verification tasks since they learn similarities/distances between a pair of inputs. As illustrated in Fig. 7, each

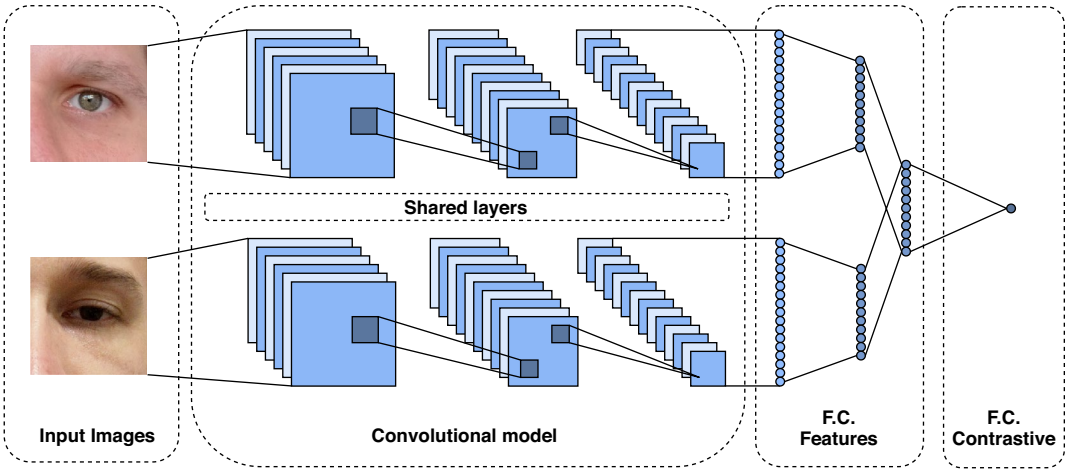


Figure 7. Siamese CNN architecture. This model is composed of two twin branches of convolutional layers sharing their trainable parameters. The output computes a distance between the input image pairs.

#	Layer	Connected to	Input	Output
0	Branch_a (MobileNetV2 (88 layers))	–	224 × 224 × 3	256
1	Branch_b (MobileNetV2 (88 layers))	–	224 × 224 × 3	256
2	Dense	#0 and #1	512	256
3	Euclidean dist. / Contrastive loss	#2	256	1

Table 4. Siamese network architecture description.

branch of the Siamese structure is composed of a CNN model followed by some dense layers. These models can also have shared and non-shared dense layers at the top.

As detailed in Table 4, we employ MobileNetV2 as the base model for each branch of the Siamese network. We use the contrastive loss[96,97] in the training stage to compute the similarity between the input pair images.

As described in[97], let D_W be the Euclidean distance between two input vectors, the contrastive loss can be written as follows:

$$C(W) = \sum_{i=1}^{P} L(W, (Y, \mathbf{X_1}, \mathbf{X_2})^i), \tag{1}$$

where

$$L(W, (Y, \mathbf{X_1}, \mathbf{X_2})^i) = (1 - Y)L_S(D_W^i) + YL_D(D_W^i), \tag{2}$$

and P is the number of training pairs, $(Y, \mathbf{X_1}, \mathbf{X_2})^i$ corresponds to the i-th label (Y) of the sample pair $\mathbf{X_1}, \mathbf{X_2}$, and L_S and L_D are partial losses for a pair of similar and dissimilar points, respectively. The objective of this function is to minimize L for L_S and L_D by computing low and high values of D_W for similar and dissimilar pairs, respectively.

The contrastive loss was proposed and applied to face verification[96,97] and has been employed for periocular recognition[98,99] and iris recognition[69].

Results and discussion

This section presents the benchmark results for the identification and verification tasks. We first describe the experimental setup used to perform the benchmark. Then, we report and discuss the results achieved by each approach.

Experimental setup. Inspired by several recent works[6,32,34,35,37,63,69,82,100], we perform the benchmark employing pre-trained models on ImageNet and also for face recognition (VGG16-Face and ResNet50-Face). Afterward, we fine-tuned these models using the UFPR-Periocular dataset. Similar to recent works on ocular recognition[7,32,35,36], we modify all models by adding a fully convolutional layer before the last layer (softmax) to generate a feature vector with a size of 256 for each image. The default input size of the models is 224 × 224 × 3, except for the InceptionResNet and Xception models, which have an input size of 299 × 299 × 3. Note that the input dimensions are different because we are using pre-trained models and therefore our fine-tuning process should follow the original architectures' input size. In this way, for training and evaluation, the periocular images were resized to fit the input size required for each method, i.e., 299 × 299 × 3 for both InceptionResNet and Xception and 244 × 244 × 3 for the remaining models.

Model	Size (MB)	Trainable parameters
VGG16	1088	135, 886, 084
VGG16-Face	1088	135, 886, 084
InceptionResNet	445	55, 246, 372
ResNet50V2	400	49, 786, 436
ResNet50	198	24, 609, 284
ResNet50-Face	198	24, 609, 284
Xception	176	21, 908, 204
DenseNet121	64	7, 792, 964
MobileNetV2	26	3, 128, 516
Multi-task	37	4, 494, 230
Siamese	21	2, 551, 808
Pairwise	20	2, 349, 479

Table 5. Size (MB) and number of trainable parameters of the CNN models used in the benchmark.

For all methods, the training was performed during 60 epochs with a learning rate of 10^{-3} for the first 15 epochs and 5×10^{-4} for the remaining epochs using the Stochastic Gradient Descent (SGD) optimizer. Then, we used the weights from the epoch that achieves the lower loss in the validation set to perform the evaluation.

We employ Rank 1 and Rank 5 accuracy for the identification task, and the Area Under the Curve (AUC), Equal Error Rate (EER), and Decidability (DEC) metrics for verification. Furthermore, to generate the verification scores, we compute the cosine distance between the deep representations generated by each CNN model. As described and applied in several works with state-of-the-art results[5,6,32,35], the cosine distance is computed by the cosine angle between two vectors, being invariant to scalar transformation. This measure gives more attention to the orientation than to the coefficient of magnitude of the representations, being an interesting metric to compute the similarity between two vectors. The cosine metric distance is given by:

$$d_c(A, B) = 1 - \frac{\sum_{j=1}^{N} A_j B_j}{\sqrt{\sum_{j=1}^{N} A_j^2} \sqrt{\sum_{j=1}^{N} B_j^2}}, \tag{3}$$

where A and B stand for the feature vectors.

Regarding the models explicitly developed for the verification tasks, i.e., the Siamese and the Pairwise Filters networks, as this task has unbalanced samples of genuine and impostors pairs, selecting the best samples to perform the training is challenging. Thus, trying to fit the models by feeding them samples as diverse as possible, we employed all genuine pairs and randomly selected the same number from the impostor pairs for each epoch. Hence, each epoch may have different impostor samples. However, for a fair comparison, we generated the random impostor pairs only once for each epoch and fold, and used the same samples for training both models.

The reported results are from five repetitions for each fold, except for the Siamese and Pairwise filter networks, in which we ran only three repetitions due to the high computational cost. All experiments were performed on a computer with an AMD Ryzen Threadripper 1920X 3.5GHz (4.0GHz Turbo) CPU, 64 GB of RAM and an NVIDIA Quadro RTX 8000 GPU (48 GB). All CNN models were implemented in Python using the Tensorflow (https://www.tensorflow.org/) and Keras (https://keras.io/) frameworks.

Benchmark results. The results obtained by each approach in the closed-world and open-world protocols are presented in this section. An ablation study were performed evaluating each task's influence in the identification mode on the Multi-task learning network. Table 5 shows the size and the number of trainable parameters of each CNN model used as a benchmark. This information was extracted from the models employed in the closed-world protocol since they have more neurons on the last layer than the open-world protocol models. We also report the results achieved by employing the state-of-the-art method that achieved first place in the VISOB 2 competition on mobile ocular biometric recognition[101]. This method[6] consists of an ensemble of five ResNet-50 models pre-trained for face recognition and fine-tuned using the periocular images of our dataset and employing the same experimental protocol described in this work.

As can be seen, the benchmark has a great diversity of models with different sizes and parameters due to their difference in structure, depth, concept, and architectures.

Closed-world protocol. We perform the benchmark for both the identification and verification tasks in the closed-world protocol. All results are presented in Table 6 and Fig. 8. Even though the MobileNetV2 is the shortest model in size and trainable parameters, it achieved the best results for identification and verification tasks. Therefore, we employed MobileNetV2 as the base model for the Multi-task, Siamese, and Pairwise Filters networks.

The Multi-task model achieved the best results in Rank 1, Rank 5, AUC, and EER metrics. We emphasize that we only explored other tasks such as—age, gender, eye side, and mobile device model—at the training stage of this model. We extracted the representations only for the classification task to evaluate identification (using the

Model	Identification (1:*N*)		Verification (1:1)		
	Rank 1 (%)	Rank 5 (%)	AUC (%)	EER (%)	Decidability
VGG16	50.56 ± 3.30	68.73 ± 3.01	99.41 ± 0.11	3.59 ± 0.32	4.4544 ± 0.1502
VGG16-Face	56.29 ± 1.62	73.84 ± 1.48	99.43 ± 0.08	3.44 ± 0.28	4.5069 ± 0.1379
Xception	57.43 ± 1.43	75.88 ± 1.52	99.77 ± 0.04	2.19 ± 0.18	4.2470 ± 0.0538
ResNet50V2	63.18 ± 2.14	77.79 ± 1.81	99.74 ± 0.04	2.24 ± 0.18	4.9382 ± 0.1184
InceptionResNet	65.16 ± 2.45	81.53 ± 1.99	99.78 ± 0.15	1.85 ± 0.40	4.5561 ± 0.1183
ResNet50	71.06 ± 1.14	85.22 ± 0.82	99.89 ± 0.02	1.41 ± 0.10	5.1242 ± 0.0634
ResNet50-Face	73.76 ± 1.43	86.86 ± 1.02	99.83 ± 0.03	1.74 ± 0.12	5.2400 ± 0.0837
DenseNet121	75.54 ± 1.36	88.53 ± 0.97	99.93 ± 0.02	1.11 ± 0.09	5.1730 ± 0.0497
MobileNetV2	77.98 ± 1.08	90.19 ± 0.79	99.93 ± 0.01	1.13 ± 0.07	5.2477 ± 0.0650
Multi-task	**84.32 ± 0.71**	**94.55 ± 0.58**	**99.96 ± 0.01**	**0.81 ± 0.06**	5.1978 ± 0.0340
Visob 2.0 Winner[6,101]	–	–	99.94 ± 0.01	1.02 ± 0.09	6.0345 ± 0.0788
Siamese	–	–	98.94 ± 0.22	4.86 ± 0.44	3.0005 ± 0.1871
Pairwise	–	–	99.44 ± 0.66	3.06 ± 1.84	**6.4503 ± 1.2270**

Table 6. Benchmark results in the closed-world protocol for the identification and verification tasks. Significant values are given in bold.

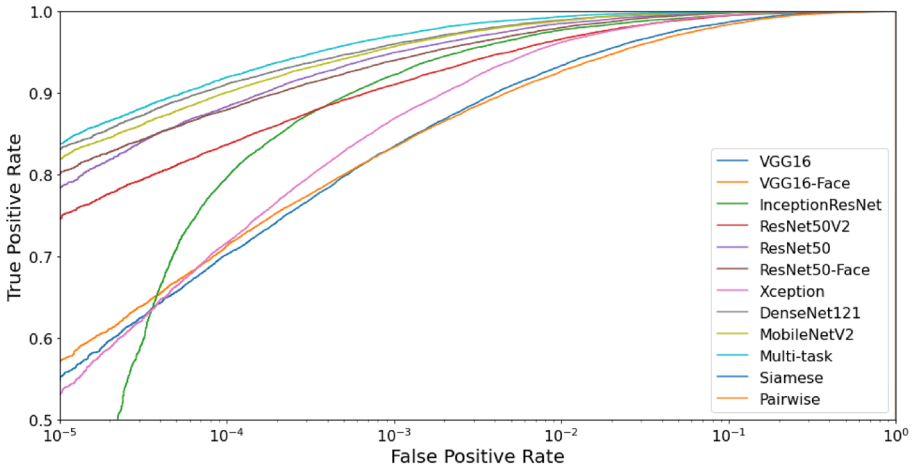


Figure 8. Receiver operating characteristic curve to compare methods in the closed-world protocol.

softmax layer) and verification (using the cosine distance) tasks. The Siamese network obtained the worst results in the benchmark. In contrast, the Pairwise Filters network reached the higher Decidability index, indicating that it was the most useful to separate genuine and impostors distributions. Nevertheless, it did not achieve the best results in terms of AUC and EER.

The models pre-trained for face recognition generally achieve best results than those pre-trained on the ImageNet dataset as stated in some previous works[32,100].

Open-world protocol. The main idea of the open-world protocol is to evaluate the capability of the methods to extract discriminant features from samples of classes that are not present in the training stage. Thus, for this protocol, we perform a benchmark only for the verification task. The results are shown in Table 7 and Fig. 9.

As in the closed-world protocol, the Multi-task model achieved the best results in Rank 1, Rank 5, AUC, and EER, and the Pairwise network achieved the best Decidability index. The Siamese and Pairwise Filters networks trained using the closed-world validation split reached better results than when trained using the open-world validation split. We believe this occurred due to the fact that there are fewer classes in the training set in the open-world validation split than in the closed-world validation split. Although the open-world validation split corresponds to a more realistic scenario regarding the test set, the networks trained with samples from a larger number of classes can reach a higher capability of generalization, producing discriminative representations even for samples from classes not present in the training stage.

Multi-task learning. The Multi-task model reached the best results in the closed- and open-world protocols. As this network simultaneously learns different tasks, we perform an ablation study by running some experiments with 4 new models created by removing one of the tasks at a time. The experiments were carried out in

Model	Validation	Verification (1:1)		
		AUC (%)	EER (%)	Decidability
VGG16	Closed-World	97.38 ± 0.53	8.52 ± 0.92	2.9599 ± 0.1572
VGG16-Face	Closed-World	97.70 ± 0.42	7.78 ± 0.75	3.0327 ± 0.1428
ResNet50	Closed-World	98.60 ± 0.28	5.98 ± 0.67	3.3702 ± 0.1413
ResNet50V2	Closed-World	98.73 ± 0.28	5.69 ± 0.64	3.4312 ± 0.1459
Xception	Closed-World	98.93 ± 0.16	5.23 ± 0.42	3.3493 ± 0.0712
InceptionResNet	Closed-World	99.10 ± 0.24	4.61 ± 0.65	3.4982 ± 0.1208
ResNet50-Face	Closed-World	99.18 ± 0.16	4.38 ± 0.47	3.8319 ± 0.1239
DenseNet121	Closed-World	99.51 ± 0.12	3.39 ± 0.46	3.8646 ± 0.1215
MobileNet	Closed-World	99.56 ± 0.08	3.17 ± 0.33	3.9868 ± 0.1067
Multi-task	**Closed-World**	**99.67 ± 0.08**	**2.81 ± 0.39**	3.9263 ± 0.0921
Visob 2.0 Winner[6,101]	-	99.65 ± 0.09	2.96 ± 0.26	4.3666 ± 0.1453
Siamese	Closed-World	97.27 ± 0.64	8.10 ± 1.01	2.6678 ± 0.2433
Pairwise	Closed-World	98.62 ± 0.72	5.77 ± 1.57	**4.4404 ± 0.5834**
Siamese	Open-World	96.85 ± 0.70	8.87 ± 1.14	2.6218 ± 0.1514
Pairwise	Open-World	97.80 ± 2.03	7.11 ± 3.66	4.1977 ± 1.0663

Table 7. Benchmark results in the open-world protocol for the verification task. Significant values are given in bold.

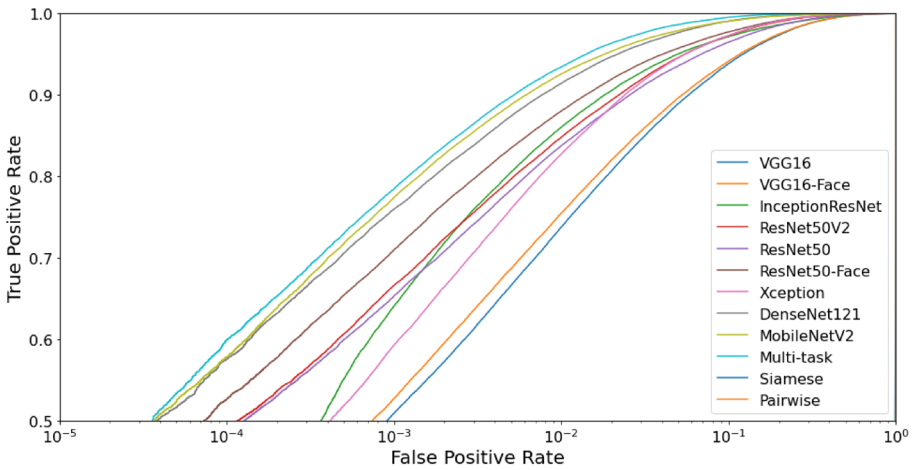


Figure 9. Receiver operating characteristic curve to compare methods in the open-world protocol.

Model	Rank 1	Rank 5	Device model	Age	Gender	Eye side
Multi-task (no model)	80.76 ± 0.94	91.96 ± 0.51	–	82.14 ± 0.83	97.72 ± 0.17	**99.99 ± 0.01**
Multi-task (no age)	81.93 ± 0.99	93.51 ± 0.69	87.20 ± 0.63	–	97.65 ± 0.20	**99.99 ± 0.01**
Multi-task (no gender)	82.48 ± 0.64	93.55 ± 0.52	86.71 ± 0.54	83.17 ± 0.54	–	**99.99 ± 0.01**
Multi-task (no side)	83.72 ± 0.61	94.07 ± 0.54	87.22 ± 0.79	83.75 ± 0.53	97.70 ± 0.20	–
Multi-task	**84.32 ± 0.71**	**94.55 ± 0.58**	**87.42 ± 0.65**	**84.34 ± 0.71**	**97.80 ± 0.21**	99.98 ± 0.02

Table 8. Results (%) from several multi-task models trained to predict different tasks. The device model concerns the task of identifying the smartphone model with which the image was taken. The age, gender, and eye side regard the tasks of classifying the input image into age ranges, gender (male or female), and eye side (left or right), respectively. Significant values are given in bold.

the closed-world protocol evaluating the performance for identification and verification. We also evaluated the results achieved by all models in each task.

According to Table 8, the Multi-task network without the prediction of the mobile device model was the most penalized for the identification task, followed by the network variations without age, gender, and eye side estimation, respectively. All models handled the gender and eye side classification tasks well, while the device

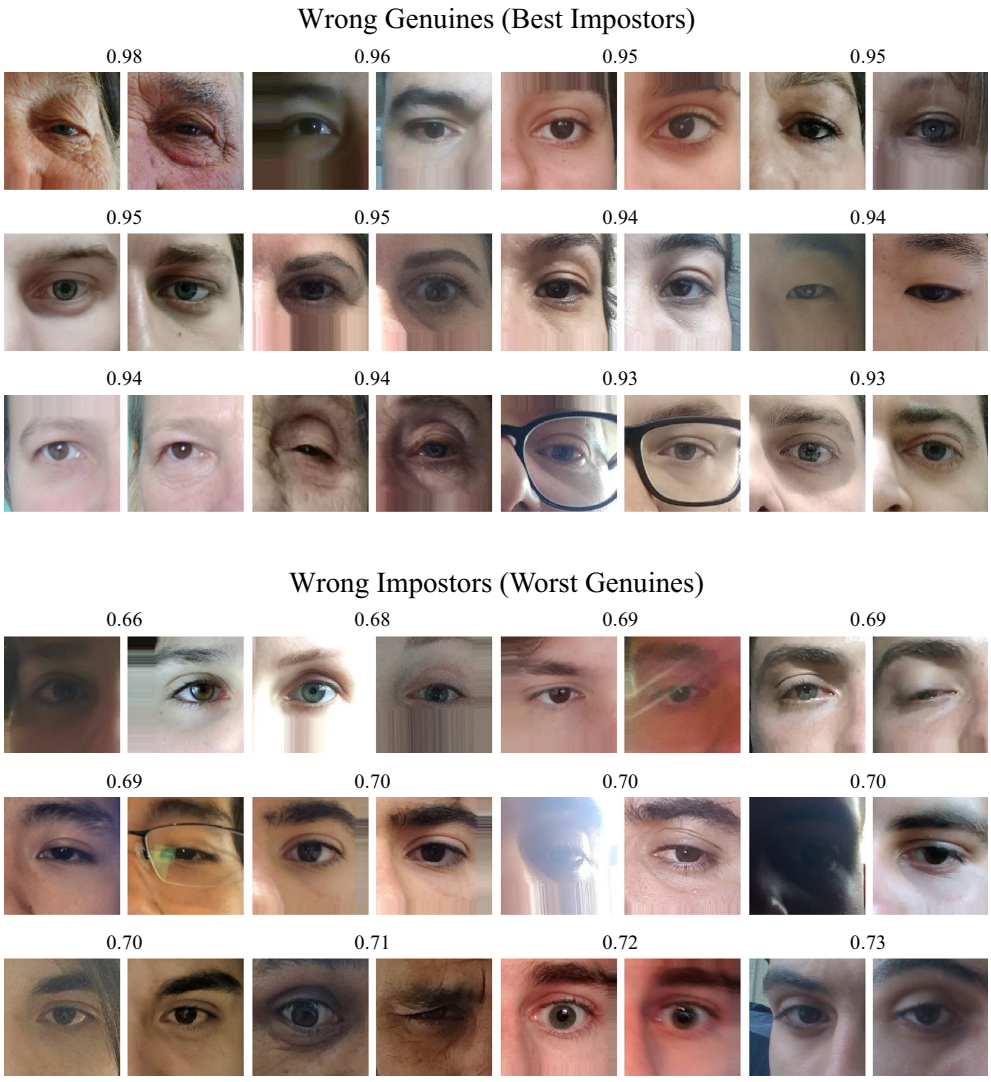


Figure 10. Pairwise images wrongly classified by the model that obtained the best result in the verification task in the open-world protocol. Higher scores mean that the pair of periocular images is more likely to be genuine.

model and age range classification tasks proved to be more challenging. One problem in the device model and age range classification is the unbalanced number of samples per class. Such bias probably contributed to the lower results being achieved in these two tasks.

Note that we only employed the class prediction for the matching in both closed-world and open-world protocols. However, as shown in Table 8, the multi-task architecture also achieved promising results in the other tasks. In this sense, it may be possible to further improve the recognition results by adopting heuristic rules based on the scores of the other tasks.

Subjective evaluation. In this section, we perform a subjective evaluation through visual inspection on the pairs of images erroneously classified by the Multi-task model, which achieved the best result in the verification task in the closed-world protocol. The best impostors (impostors classified as genuine) and the worst genuines (genuine classified as impostors) pairs are presented in Fig. 10.

Performing a visual analysis of all pairwise errors, it is clear that hair occlusion, age, eyeglasses, and eye shape were the most influential factors that led the model to the wrong classification of genuine pairs (intra-class comparison). In pairs wrongly classified as impostors (inter-class comparison), we saw that lighting, blur, eyeglasses, off-angle, eye-gaze, reflection, and facial expression caused the main difference between the images. We hypothesize that some errors caused by lightning, blur, reflection, and occlusion can be reduced by employing some data augmentation techniques in the training stage. Attribute normalization[5] can also reduce the errors caused by attributes present in the periocular region such as eyeglasses, eye gaze, makeup, and some types of occlusion. Although some methods can be applied to reduce the matching errors, there are still several characteristics in these images that make the mobile periocular recognition a challenging task, mainly to the high intra-class variations.

Conclusion

This article introduces a new periocular dataset that contains images captured in unconstrained environments on different sessions using several mobile device models. The main idea was to create a dataset with real-world images regarding lighting, noises, and attributes in the periocular region. To the best of our knowledge, in the literature, this is the first periocular dataset with more than 1, 000 subject samples and the largest one in the number of different sensors (196).

We presented an extensive benchmark with several CNN models and architectures employed in recent works for periocular recognition. These architectures consist of models for multi-class classification and multi-task learning, in addition to Siamese and pairwise filters networks. We evaluated the methods in the closed-world and open-world protocols, as well as for the identification and verification tasks. For both protocols and tasks, the multi-task model achieved the best results. Thus, we conducted an ablation study on this model to understand which tasks significantly influenced the results. We stated that the mobile device model identification task was the most important, followed by age range, gender, and eye side classification. Note that we did not conduct experiments employing only left or right eye sides or images separated by gender. The model trained using all these tasks reported the best result for the identification and verification in the closed- and open-world protocols.

In a complementary way, we performed a subjective analysis of the best/worst false genuine and true impostors image pairwise comparisons using the Multi-task model, which achieved the best performance for the verification task. We observed that lighting, occlusion, and image resolution were the most critical factors that led the model to wrong verification.

We believe that the UFPR-Periocular dataset will be of great relevance to assist in evolving periocular biometric systems using images obtained by mobile devices in unconstrained scenarios. This dataset is the most extensive in terms of the number of subjects in the literature and has natural within-class variability due to samples captured in different sessions.

The Multi-task network using MobileNetV2 as baseline model achieved the best benchmark results for the identification and verification tasks, reaching a rank 1 of 84, 32% and an EER of 0.81% in the closed-world protocol, and an EER of 2.81% in the open-world protocol with thresholds of 0.80 and 0.78, respectively. Therefore, there is still room for improvement in both identification and verification tasks.

Data availability

The UFPR-Periocular dataset is publicly available for the research community (upon request) at https://web.inf.ufpr.br/vri/databases/ufpr-periocular/. The dataset contains all the original and cropped periocular images, along with the eye corner annotations we made manually. The files determining all splits and setups for training, validation, and testing employed in our experiments are also part of the dataset as well as information about age, gender, and device model for each image. We recognize the importance of also providing mask labels for sclera and iris segmentation; however, this is left for future work as making such annotations is very time-consuming[29,30]. The UFPR-Periocular dataset is released only to academic researchers from educational or research institutes for non-commercial purposes. To be able to download the dataset, please read carefully the license agreement available at https://web.inf.ufpr.br/vri/wp-content/uploads/sites/7/2020/11/UFPR-Periocular-License-Agreement.pdf, fill it out, and send it back to Professor David Menotti (*menotti@inf.ufpr.br*). The license agreement must be reviewed and signed by the individual or entity authorized to make legal commitments on behalf of the institution or corporation (e.g., Department/Administrative Head, or similar).

Received: 15 May 2022; Accepted: 19 October 2022
Published online: 26 October 2022

References

1. Santos, G. & Proença, H. Periocular biometrics: An emerging technology for unconstrained scenarios. In 2013 IEEE Symposium on Computational Intelligence in Biometrics and Identity Management (CIBIM), 14–21. https://doi.org/10.1109/CIBIM.2013.6607908 (2013).
2. De Marsico, M., Nappi, M. & Proença, H. Results from MICHE II - Mobile Iris CHallenge Evaluation II. *Pattern Recogn. Lett.* **91**, 3–10 (2017).
3. Proença, H. & Neves, J. C. IRINA: Iris recognition (even) in inaccurately segmented data. *IEEE Conf. Comput. Vis. Patt. Recognit. (CVPR)* **1**, 6747–6756 (2017).
4. Proença, H. & Neves, J. C. A reminiscence of "mastermind": Iris/periocular biometrics by "in-set" CNN iterative analysis. *IEEE Trans. Inf. Foren. Secur.* **14**, 1702–1712 (2019).
5. Zanlorensi, L. A., Proença, H. & Menotti, D. Unconstrained periocular recognition: Using generative deep learning frameworks for attribute normalization. In *2020 International Conference on Image Processing (ICIP)*, 1361–1365 (2020).
6. Zanlorensi, L. A., Lucio, D. R., Britto, A. S. Jr., Proença, H. & Menotti, D. Deep representations for cross-spectral ocular biometrics. *IET Biometrics* **9**, 68–77 (2020).
7. Zanlorensi, L. A. *et al.* Ocular recognition databases and competitions: A survey. *Artif. Intell. Rev.* **55**, 129–180 (2022).
8. Zheng, W.-S., Gong, S. & Xiang, T. Towards open-world person re-identification by one-shot group-based verification. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**, 591–606 (2016).
9. Leng, Q., Ye, M. & Tian, Q. A survey of open-world person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **30**, 1092–1108 (2020).
10. Raja, K. B., Raghavendra, R., Vemuri, V. K. & Busch, C. Smartphone based visible iris recognition using deep sparse filtering. *Pattern Recogn. Lett.* **57**, 33–42 (2015).
11. Santos, G., Grancho, E., Bernardo, M. V. & Fiadeiro, P. T. Fusing iris and periocular information for cross-sensor recognition. *Pattern Recogn. Lett.* **57**, 52–59 (2015).
12. Algashaam, F. M. *et al.* Multispectral periocular classification with multimodal compact multi-linear pooling. *IEEE Access* **5**, 14572–14578 (2017).
13. Sharma, A., Verma, S., Vatsa, M. & Singh, R. On cross spectral periocular recognition. In IEEE International Conference on Image Processing (ICIP), 5007–5011 (2014).

14. Dobeš, M., Machala, L., Tichavský, P. & Pospíšil, J. Human eye iris recognition using the mutual information. *Optik - Int. J. Light Electron Opt.* **115**, 399–404 (2004).
15. Hosseini, M. S., Araabi, B. N. & Soltanian-Zadeh, H. Pigment melanin: Pattern for iris recognition. *IEEE Trans. Instrum. Meas.* **59**, 792–804 (2010).
16. De Marsico, M., Nappi, M., Riccio, D. & Wechsler, H. Mobile Iris Challenge Evaluation (MICHE)-I, biometric iris dataset and protocols. *Pattern Recogn. Lett.* **57**, 17–23 (2015).
17. Sequeira, A. *et al.* Cross-eyed—Cross-spectral iris/periocular recognition database and competition. *Int. Conf. Biometr. Spec. Int. Group* **260**, 1–5 (2016).
18. Sequeira, A. F. et al. Cross-Eyed 2017: Cross-spectral iris/periocular recognition competition. In *IEEE International Joint Conference on Biometrics*, 725–732 (2017).
19. Nalla, P. R. & Kumar, A. Toward more accurate iris recognition using cross-spectral matching. *IEEE Trans. Image Process.* **26**, 208–221 (2017).
20. Proença, H. & Alexandre, L. A. UBIRIS: A noisy iris image database. In *Image Analysis and Processing (ICIAP)*, 970–977 (2005).
21. Proença, H., Filipe, S., Santos, R., Oliveira, J. & Alexandre, L. A. The UBIRIS.v2: A database of visible wavelength iris images captured on-the-move and at-a-distance. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 1529–1535 (2010).
22. Padole, C. N. & Proença, H. Periocular recognition: Analysis of performance degradation factors. In *IAPR International Conference on Biometrics (ICB)*, 439–445 (2012).
23. Rattani, A., Derakhshani, R., Saripalle, S. K. & Gottemukkula, V. ICIP 2016 competition on mobile ocular biometric recognition. In *IEEE International Conference on Image Processing – Challenge Session on Mobile Ocular Biometric Recognition*, 320–324 (2016).
24. Menotti, D. *et al.* Deep representations for iris, face, and fingerprint spoofing detection. *IEEE Trans. Inf. Forensics Secur.* **10**, 864–879 (2015).
25. He, L. *et al.Multi-patch convolution neural network for iris liveness detection* 1–7. In *IEEE International Conf. on Biometrics Theory, Applications and Systems* (2016).
26. Silva, P. et al. An approach to iris contact lens detection based on deep image representations. In *Conference on Graphics, Patterns and Images (SIBGRAPI)*, 157–164 (2015).
27. Lucio, D. R., Laroca, R., Zanlorensi, L. A., Moreira, G. & Menotti, D. Simultaneous iris and periocular region detection using coarse annotations. In *Conference on Graphics, Patterns and Images (SIBGRAPI)*, 178–185 (2019).
28. Severo, E. et al. A benchmark for iris location and a deep learning detector evaluation. In *International Joint Conference on Neural Networks (IJCNN)*, 1–7 (2018).
29. Lucio, D. R., Laroca, R., Severo, E., Britto Jr., A. S. & Menotti, D. Fully convolutional networks and generative adversarial networks applied to sclera segmentation. In *IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 1–7 (2018).
30. Bezerra, C. S. et al. Robust iris segmentation based on fully convolutional networks and generative adversarial networks. In *Conference on Graphics, Patterns and Images*, 281–288 (2018).
31. Du, Y., Bourlai, T. & Dawson, J. Automated classification of mislabeled near-infrared left and right iris images using convolutional neural networks. *BTAS* 1–6 (2016).
32. Luz, E., Moreira, G., Zanlorensi Junior, L. A. & Menotti, D. Deep periocular representation aiming video surveillance. *Pattern Recognit. Lett.* **114**, 2–12 (2018).
33. Zhao, T., Liu, Y., Huo, G. & Zhu, X. A deep learning iris recognition method based on capsule network architecture. *IEEE Access* **7**, 49691–49701 (2019).
34. Diaz, K. H., Alonso-Fernandez, F. & Bigun, J. Spectrum translation for cross-spectral ocular matching. arXiv preprint arXiv: 2002.06228 (2020).
35. Zanlorensi, L. A. et al. The impact of preprocessing on deep representations for iris recognition on unconstrained environments. In *Conference on Graphics, Patterns and Images (SIBGRAPI)*, 289–296 (2018).
36. Silva, P. H. et al. Multimodal feature level fusion based on particle swarm optimization with deep transfer learning. In *2018 Congress on Evolutionary Computation (CEC)*, 1–8 (2018).
37. Hernandez-Diaz, K., Alonso-Fernandez, F. & Bigun, J. Cross-spectral periocular recognition with conditional adversarial networks. In *IEEE International Joint Conference on Biometrics (IJCB)*, 1–9 (2020).
38. Proença, H. & Alexandre, L. A. Toward covert iris biometric recognition: Experimental results from the NICE contests. *IEEE Trans. Inf. Forensics Secur.* **7**, 798–808 (2012).
39. Rattani, A., Derakhshani, R. & Ross, A. *Introduction to Selfie Biometrics, 1–18* (Springer International Publishing, Cham, 2019).
40. Tapia, J. E., Valenzuela, A., Lara, R., Gomez-Barrero, M. & Busch, C. Selfie periocular verification using an efficient super-resolution approach. *IEEE Access* **10**, 67573–67589. https://doi.org/10.1109/ACCESS.2022.3184301 (2022).
41. Alonso-Fernandez, F., Farrugia, R. A., Fierrez, J. & Bigun, J. *Super-resolution for Selfie Biometrics: Introduction and Application to Face and Iris, 105–128* (Springer International Publishing, Cham, 2019).
42. Khellat-Kihel, S., Lagorio, A. & Tistarelli, M. *Foveated Vision for Biologically Inspired Continuous Face Authentication, 129–143* (Springer International Publishing, Cham, 2019).
43. Arora, G., Tiwari, K. & Gupta, P. *Liveness and Threat Aware Selfie Face Recognition, 197–210* (Springer International Publishing, Cham, 2019).
44. Vitek, M. et al. Ssbc 2020: Sclera segmentation benchmarking competition in the mobile environment. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, 1–10 (2020).
45. Phillips, P. J., Bowyer, K. W., Flynn, P. J., Liu, X. & Scruggs, W. T. The iris challenge evaluation 2005. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, 1–8 (2008).
46. Phillips, P. J. *et al.* FRVT 2006 and ICE 2006 large-scale experimental results. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 831–846 (2010).
47. Shah, S. & Ross, A. Generating synthetic irises by feature agglomeration. In *International Conf. on Image Processing*, 317–320 (2006).
48. Zuo, J., Schmid, N. A. & Chen, X. On generation and analysis of synthetic iris images. *IEEE Trans. Inf. Forensics Secur.* **2**, 77–90 (2007).
49. Ruiz-Albacete, V., Tome-Gonzalez, P., Alonso-Fernandez, F., Galbally, J. & Ortega-Garcia, J. Direct attacks using fake images in iris verification. In *Biometrics and Identity Management*, 181–190 (2008).
50. Czajka, A. Database of iris printouts and its application: Development of liveness detection method for iris recognition. In *International Conf. on Methods Models in Automation Robotics*, 28–33 (2013).
51. Gupta, P., Behera, S., Vatsa, M. & Singh, R. On iris spoofing using print attack. In *International Conference on Pattern Recognition (ICPR)*, 1681–1686 (2014).
52. Kohli, N., Yadav, D., Vatsa, M., Singh, R. & Noore, A. *Detecting medley of iris spoofing attacks using DESIST* 1–6 (In IEEE Intl. Conf. on Biometrics Theory, Applications and Systems, 2016).
53. Baker, S. E., Hentz, A., Bowyer, K. W. & Flynn, P. J. Degradation of iris recognition performance due to non-cosmetic prescription contact lenses. *Comput. Vis. Image Underst.* **114**, 1030–1044 (2010).

54. Kohli, N., Yadav, D., Vatsa, M. & Singh, R. Revisiting iris recognition with color cosmetic contact lenses. *Int. Conf. Biomet. (ICB)* **1**, 1–7 (2013).
55. Doyle, J. S., Bowyer, K. W. & Flynn, P. J. Variation in accuracy of textured contact lens detection based on sensor and lens pattern. *BTAS*, 1–7 (2013).
56. Doyle, J. S. & Bowyer, K. W. Robust detection of textured contact lenses in iris recognition using BSIF. *IEEE Access* **3**, 1672–1683 (2015).
57. Fenker, S. P. & Bowyer, K. W. Analysis of template aging in iris biometrics. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 45–51 (2012).
58. Baker, S. E., Bowyer, K. W., Flynn, P. J. & Phillips, P. J. Template Aging in Iris Biometrics, chap. 11, 205–218 (Springer London, 2013).
59. Bowyer, K. W., Hollingsworth, K. & Flynn, P. J. Image understanding for iris biometrics: A survey. *Comput. Vis. Image Underst.* **110**, 281–307 (2008).
60. Proença, H. & Neves, J. C. Segmentation-less and non-holistic deep-learning frameworks for iris recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1–10 (2019).
61. Daugman, J. High confidence visual recognition of persons by a test of statistical independence. *IEEE Trans. Pattern Anal. Mach. Intell.* **15**, 1148–1161 (1993).
62. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
63. Reddy, N., Rattani, A. & Derakhshani, R. Comparison of deep learning models for biometric-based mobile user authentication. In *IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 1–6 (2018).
64. Siddiqui, H., Rattani, A., Ricanek, K. & Hill, T. An examination of bias of facial analysis based bmi prediction models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2926–2935 (2022).
65. Ramachandran, S. & Rattani, A. Deep generative views to mitigate gender classification bias across gender-race groups (2022).
66. Banerjee, A., Ghosh, C. & Mandal, S. N. Analysis of v-net architecture for iris segmentation in unconstrained scenarios. *SN Comput. Sci.* **3**, 1–24 (2022).
67. Jain, A. K. & Ross, A. *Introduction to Biometrics, 1–22* (Springer, US, 2008).
68. Deng, J. et al. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (2009).
69. Wang, K. & Kumar, A. Cross-spectral iris recognition using cnn and supervised discrete hashing. *Pattern Recogn.* **86**, 85–98 (2019).
70. Platt, J. C., Cristianini, N. & Shawe-Taylor, J. Large margin dags for multiclass classification. In *International Conference on Neural Information Processing Systems (NIPS)* (1999).
71. Hastie, T., Rosset, S., Zhu, J. & Zou, H. Multi-class adaboost. *Stat. Interface* **2**, 349–360 (2009).
72. Huang, G., Zhou, H., Ding, X. & Zhang, R. Extreme learning machine for regression and multiclass classification. *IEEE Trans. Syst. Man Cybern Part B (Cybernetics)* **42**, 513–529 (2012).
73. Zhang, Q., Li, H., Sun, Z., He, Z. & Tan, T. Exploring complementary features for iris recognition on mobile devices. In *2016 International Conference on Biometrics (ICB)*, 1–8 (2016).
74. Donida Labati, R., Genovese, A., Piuri, V., Scotti, F. & Vishwakarma, S. I-social-db: A labeled database of images collected from websites and social media for iris recognition. *Image Vis. Comput.* **105**, 1–9. (2021).
75. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)* (2015).
76. Parkhi, O. M., Vedaldi, A. & Zisserman, A. Deep face recognition. In *British Machine Vision Conference (BMVC)*, 1–12 (2015).
77. Zhao, T., Liu, Y., Huo, G. & Zhu, X. A deep learning iris recognition method based on capsule network architecture. *IEEE Access* **7**, 49691–49701 (2019).
78. Behera, S. S., Mishra, S. S., Mandal, B. & Puhan, N. B. Variance-guided attention-based twin deep network for cross-spectral periocular recognition. *Image Vis. Comput.* 104016 (2020).
79. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778 (2016).
80. Cao, Q., Shen, L., Xie, W., Parkhi, O. M. & Zisserman, A. VGGFace2: A dataset for recognising faces across pose and age. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 67–74 (2018).
81. Boyd, A., Czajka, A. & Bowyer, K. *Deep learning-based feature extraction in iris recognition: Use existing models, fine-tune or train from scratch?* 1–9 (In IEEE International Conf. on Biometrics Theory, Applications and Systems, 2019).
82. Boutros, F. et al. Fusing iris and periocular region for user verification in head mounted displays. In *IEEE International Conference on Information Fusion (FUSION)*, 1–8 (2020).
83. He, K., Zhang, X., Ren, S. & Sun, J. Identity mappings in deep residual networks. In *European Conf. on Computer Vision*, 630–645 (2016).
84. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *ICLR 2016 Workshop* (2016).
85. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826 (2016).
86. Szegedy, C. et al. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9 (2015).
87. Howard, A. G. et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017).
88. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L. MobileNetV2: Inverted residuals and linear bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition* (2018).
89. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).
90. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).
91. Caruana, R. Multitask learning. *Mach. Learn.* **28**, 41–75 (1997).
92. Ruder, S. An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017).
93. Laroca, R., Araujo, A. B., Zanlorensi, L. A., De Almeida, E. C. & Menotti, D. Towards image-based automatic meter reading in unconstrained scenarios: A robust and efficient approach. *IEEE Access* **9**, 67569–67584 (2021).
94. Liu, N., Zhang, M., Li, H., Sun, Z. & Tan, T. DeepIris: Learning pairwise filter bank for heterogeneous iris verification. *Pattern Recogn. Lett.* **82**, 154–161 (2016).
95. Bromley, J., Guyon, I., LeCun, Y., .Säckinger, E. & Shah, R. Signature verification using a "Siamese" time delay neural network. In *Intl. Conf. on Neural Information Processing Systems,* 737–744 (1993).
96. Chopra, S., Hadsell, R. & LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. *IEEE Conf. Comput. Vis. Pattern Recognit.* **1**, 539–546 (2005).
97. Hadsell, R., Chopra, S. & LeCun, Y. Dimensionality reduction by learning an invariant mapping. *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)* **2**, 1735–1742 (2006).

98. Zhao, Z. & Kumar, A. Improving periocular recognition by explicit attention to critical regions in deep neural network. *IEEE Trans. Inf. Forensics Secur.* **13**, 2937–2952 (2018).
99. Behera, S. S., Mandal, B. & Puhan, N. B. Twin deep convolutional neural network-based cross-spectral periocular recognition. In *2020 National Conference on Communications (NCC)*, 1–6 (2020).
100. Boyd, A., Czajka, A. & Bowyer, K. Deep learning-based feature extraction in iris recognition: Use existing models, fine-tune or train from scratch? In *IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 1–9 (2019).
101. Nguyen, H. M., Reddy, N., Rattani, A. & Derakhshani, R. Visob 2.0 - the second international competition on mobile ocular biometric recognition. In Del Bimbo, A. et al. (eds.) *Pattern Recognition. ICPR International Workshops and Challenges*, 200–208 (Springer International Publishing, Cham, 2021).

Acknowledgements
This work was supported by grants from the National Council for Scientific and Technological Development (CNPq) (# 313423/2017-2 and # 428333/2016-8) and the Coordination for the Improvement of Higher Education Personnel (CAPES). We acknowledge the support of NVIDIA Corporation with the donation of the Quadro RTX 8000 GPU used for this research.

Author contributions
L.A.Z conceived the experiments, formal analysis, data curation, and original writing. L.A.Z and L.R.S conducted the software (mobile app) development and investigation. L.A.Z, R.L and D.R.L conducted the data curation, validation, investigation, and methodology. D.M and A.S.B.J. conceived the conceptualization, validation, and supervision. All authors analyzed the results and reviewed the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to L.A.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.