UFPR-Periocular: A Periocular Dataset Collected by Mobile Devices in Unconstrained Scenarios

Luiz A. Zanlorensi, Rayson Laroca, Diego R. Lucio, Lucas R. Santos, Alceu S. Britto Jr., and David Menotti

***Abstract*—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 1,122 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.**

***Index Terms*—Mobile ocular biometric, Periocular dataset, Periocular recognition, Deep representations.**

I. Introduction

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]–[5]. 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. 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.

Luiz A. Zanlorensi, Rayson Laroca, Diego R. Lucio, Lucas R. Santos, and David Menotti are with Federal University of Paraná (UFPR), Brazil. E-mails: {lazjunior, rblsantos, drlucio, lrs14, menotti}@inf.ufpr.br

Alceu S. Britto Jr. is with Pontifical Catholic University of Paraná (PUCPR), Brazil. E-mail: alceu@ppgia.pucpr.br

TABLE I
Comparison of the available ocular datasets containing visible (VIS) images with our dataset (UFPR-Periocular).

Dataset	Subjects	Images	Sessions	Sensors
VSSIRIS [6]	28	560	1	2
CSIP [7]	50	2,004	N/A	7
QUT [8]	53	212	N/A	2
IIITD [9]	62	1,240	N/A	3
UPOL [10]	64	384	N/A	1
UTIRIS [11]	79	1,540	2	2
MICHE-I [12]	92	3,732	2	3
CROSS-EYED [13], [14]	120	3,840	N/A	2
PolyU Cross-Spectral [15]	209	12,540	2	2
UBIRIS.v1 [16]	241	1,877	2	1
UBIRIS.v2 [17]	261	11,102	2	1
UBIPr [18]	261	10,950	2	1
VISOB [19]	550	**158,136**	2	3
UFPR-Periocular	**1,122**	33,660	**3**	**196**

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 [20], [21], iris and periocular region detection [22]–[24], iris and sclera segmentation [25], [26], and iris and periocular recognition [27]–[33]. 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) [17], [34], iris recognition using mobile images (MICHE.I and MICHE.II) [1], [12], iris and periocular recognition in cross-spectral scenarios (Cross-Eyed 1 and 2) [13], [14], and periocular recognition using mobile images captured in different lighting conditions (VISOB 1 and 2) [19]. 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 [1], [19]. 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 [1], [4], [35], 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 ocular 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 I, the datasets in the literature do not present a large number of subjects and have few sensors and session captures. As described in some previous works [4], [5], 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 sensors, 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. Fig. 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 biometric systems.

Fig. 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.

Note that our dataset is the largest one in terms of the number of subjects, sessions, and sensors, as shown in Table I. 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 ocular dataset with more than 1,000 subject samples and the largest one in different sensors in the literature. Thus, we believe that it can provide a new benchmark to evaluate and develop new robust ocular biometric approaches.

The remainder of this work is organized as follows. In Section II, we describe the ocular datasets containing VIS images for ocular biometrics. In Section III, we present information about the UFPR-Periocular dataset and the proposed protocol to evaluate biometric systems. Section IV presents the CNN architectures used to perform the benchmark. In Section V, we present and discuss the benchmark results. Finally, the conclusions are given in Section VI.

II. 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. [35] 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 unconstrained environments, ocular recognition on cross-spectral scenarios, iris/periocular region detection, iris/periocular region segmentation, and sclera segmentation.

Existing ocular 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 [35], datasets containing images obtained at the near-infrared (NIR) wavelength were created mainly to investigate the intricate patterns present in the iris region [36], [37]. There are also other studies on NIR ocular images, such as generating synthetic iris images [38], [39], spoofing and liveness detection [40]–[43], contact lens detection [44]–[47], and template aging [48], [49]. 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 [2], [36], [37], [50], [51].

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 [50]. 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 ocular biometric recognition using images obtained by mobile devices in uncontrolled environments using different sensors [6], [12], [19]. The following datasets were developed to investigate the use of iris and periocular traits in VIS images: UPOL [10], UBIRIS.v1 [16], UBIRIS.v2 [17] and UBIPr [18]. 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 [11], IIITD Multi-spectral Periocular [9], PolyU Cross-Spectral [15], CROSS-EYED [13], [14], and QUT Multispectral Periocular [8]. Focusing specifically on ocular recognition using non-cooperative images obtained in uncontrolled environments by mobile devices, we highlight the following datasets: MICHE-I [12], VSSIRIS [6], CSIP [7] and VISOB [19].

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 I, the most extensive dataset regarding subjects and images is VISOB [19], which has $158{,}136$ images from 550 subjects. The ICIP 2016 Competition on mobile ocular biometric recognition [19] employed this dataset, and in the WCCI/IJCNN2020 challenge[1], 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 5 ocular 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 [34], MICHE.II [1], and CROSS-EYED I [13] and II [14]. 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 [52] 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 [53], several methods proposing and applying such techniques have been developed to address different tasks using ocular images [3]–[5], [20]–[33]. Also, as found in the literature, deep learning frameworks for ocular biometric systems are a recent technology that still needs improvement [35]. 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 [1], [4], [19], [35], [54].

III. 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[2]. 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.

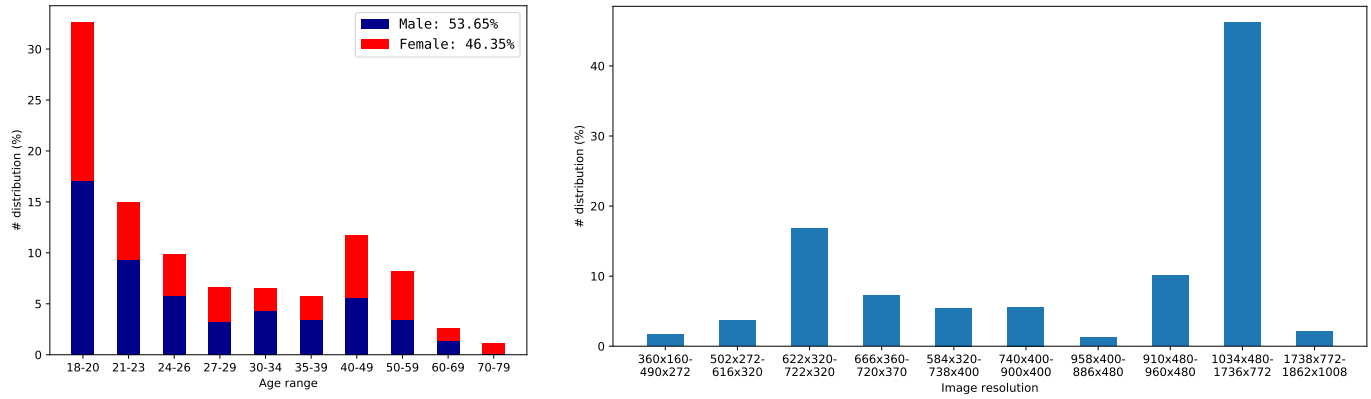
The dataset has $16{,}830$ images of both eyes from $1{,}122$ subjects. Image resolutions vary from 360×160 to 1862×1008 pixels – depending on the mobile device that was used to capture the image. We crop/separate the periocular regions of the right and left eyes to perform the benchmark, assigning a unique class to each side. Note that, once the image is cropped, the remainder image region is 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 4 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.

All the original and cropped periocular images along with the eye corner annotations are publicly available for the research community (upon request) at *https://web.inf.ufpr.br/vri/databases/ufpr-periocular/*.

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 512×512 pixels. Then, the images were split into 2 patches 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

[1]VISOB 2.0 Dataset and Competition results available at: https://sce.umkc.edu/research-sites/cibit/dataset.html.

[2]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/



(a) gender distribution among the age ranges

(b) image resolutions grouped into 10 intervals

Fig. 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.

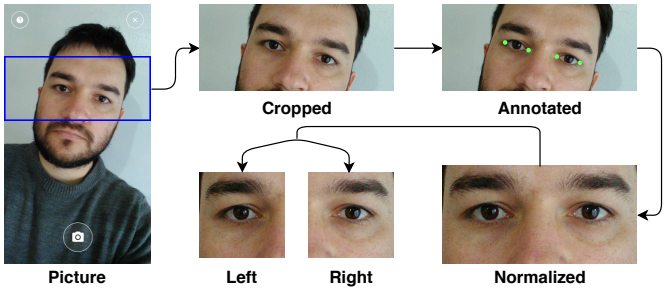


Fig. 3. Image acquisition and normalization process. First, after the subject takes the shot, the rectangular region (outlined in blue) is cropped and stored. Then, the images are normalized in terms of rotation and scale using the manual annotations of the corners of the eyes. Finally, the normalized images are cropped, generating the periocular regions of the left and right eyes.

by lighting, occlusion, specular reflection, blur, motion blur, eyeglasses, off-angle, eye-gaze, makeup, and facial expression.

A. 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 identity (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 [50]. 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 [55]. 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 II 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

TABLE II
IMAGES, CLASSES, AND PAIRWISE COMPARISON DISTRIBUTIONS FOR THE CLOSED-WORLD (CW) AND OPEN-WORLD (OW) PROTOCOLS. VALUES FOR EACH FOLD (3 FOLDS).

Protocol	Train/Val	Images / Classes			Genuine pairs / Impostor pairs		
		Train	Validation	Test	Train	Validation	Test
CW	CW/CW	13,464/2,244	8,976/2,244	11,220/2,244	33,660/ 90,599,256	13,464/40,266,336	22,440/12,583,230
OW	OW/CW	13,464/1,496	8,976/1,496	11,220/ 748	53,856/ 90,579,060	22,440/40,257,360	78,540/ 4,190,670
OW	OW/OW	15,000/1,000	7,440/ 496	11,220/ 748	105,000/112,387,500	52,080/27,621,000	78,540/ 4,190,670

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 II. The files determining all splits and setups detailed in this section are available along with the UFPR-Periocular dataset.

IV. 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 [56] and were applied in recent works of ocular recognition [5], [28], [31], [32], [57]. These methods differ from each other in network architecture, loss function, and training strategies. We employed the following CNN models: Multi-class classification, Multi-task learning, Siamese networks, and Pairwise filters networks. In the following subsections, we describe and detail each one of them.

A. 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 [58]–[60] 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 [56]. In summary, the architecture of these models has several convolutional, pooling, activation, and fully-connected layers, as shown in Fig. 4.

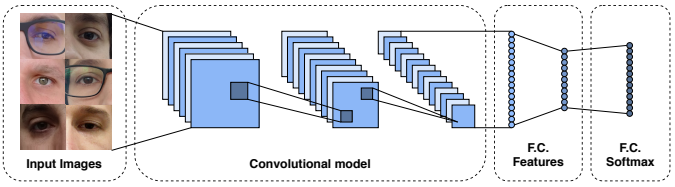


Fig. 4. Multi-class classification CNN architecture.

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. Below we describe the main characteristics of each model.

*1) **VGG**:* The VGG model, proposed by Simonyan and Zisserman [61], 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. [62] 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 [5], [28], [31], [32], [57], [63], [64]. In this

work, we employed the VGG16 and VGG16-Face to perform the benchmark.

*2) **ResNet**:* The Residual Network (ResNet) was introduced by He et al. [65] and applied to biometrics for face recognition [66], iris recognition [5], [31], [57], [63], [67] and periocular recognition [5], [33], [64], [68]. 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 [56]. In [69], 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.

*3) **InceptionResNet**:* The InceptionResNet model [70], combines the residual connections [65] and the inception architecture [71]. The first inception model [72], 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 models trained faster and reached slightly better results than the inception architecture [70]. In our experiments, we employed the InceptionResNetV2 model since it achieved the best results on ImageNet.

*4) **MobileNet**:* The first version of the MobileNet model (MobileNetV1) [73] 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 [73], 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 [73]. MobileNetV2 [74] combines the first version architecture with an inverted ResNet [65] 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 [74].

*5) **DenseNet**:* The Dense Convolutional Network (DenseNet) model [75] 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 [65], [69], 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 [56], [75]. The authors proposed different models varying the depth of the network. In our experiments, we employed DenseNet121 (the shallowest one).

*6) **Xception**:* Inception modules inspired the creation of the Xception model, which can be defined as an intermediate step between regular convolution and the depthwise separable convolution operation [76]. The proposed architecture replaces the standard inception modules with depthwise separable convolutions, and also have residual connections. The Xception architecture has the same number of parameters as InceptionV3 but outperforms it on the ImageNet dataset [56].

B. Multi-task Learning

Multi-task learning uses the domain information of related tasks as an inductive bias to improve generalization [77]. A Multi-task network can learn several tasks using a shared CNN model, where each task can help the generalization for other tasks. Caruana [77] 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 using hard or soft parameter sharing [78]. The most common one is the hard parameter sharing, where all the hidden (convolutional) layers weights are shared, i.e., the model learns a single representation for all tasks. Then, different tasks use these shared features by adding some layers for each specific task. On the other hand, in soft parameter sharing one model is employed for each task. Then, the parameters of these models are regularized to encourage similarities among them.

As shown in Fig. 4, 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.

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 III, 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.

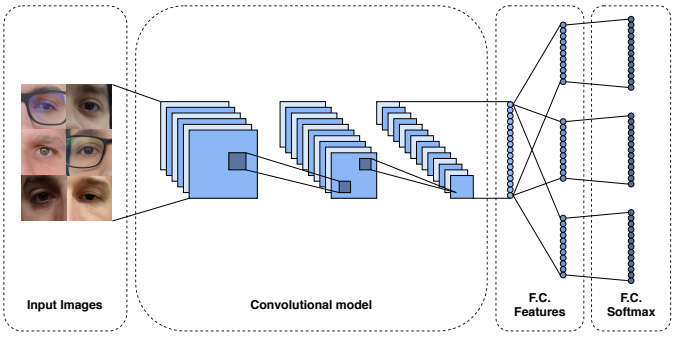


Fig. 5. Multi-task CNN architecture. In this model, each task has its own output and all tasks share the convolutional layers. The loss of all tasks is used to update the weights of the convolutional layers.

TABLE III
MULTI-TASK ARCHITECTURE IN THE CLOSED-WORLD PROTOCOL.

#	Layer	Connected to	Input	Output
0	MobileNetV2 (88 layers)	–	$224 \times 224 \times 3$	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

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 Multi-task learning networks can use weighted loss for the tasks, 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.

C. Pairwise Filters Network

Inspired by [79], 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 this model requires a pair of images as input, different concatenation strategies can be employed. Following Liu et al. [79], in this work, we generate the input pairs by concatenating the images at the depth level. Let two RGB images with shapes of $224 \times 224 \times 3$, concatenating both images by its channels; the resulting input image will have a shape of $224 \times 224 \times 6$. The output of our model has two neurons and

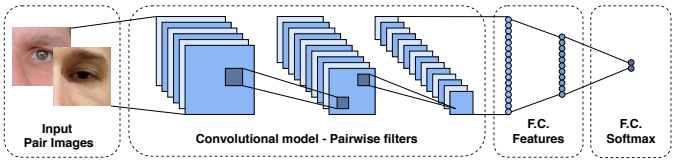


Fig. 6. Pairwise filters CNN architecture. This model contains filters that directly learn the similarity between a pair of images. The output informs whether the images are of the same person or not.

uses a softmax cross-entropy loss. As the verification problem has only two classes, this model' output can also 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.

D. Siamese Network

Introduced by Bromley et al. [80] for signature verification, Siamese networks consist of twin branches sharing their parameters (trainable parameters). Such models learn similarities/distances between a pair of inputs, being used mainly for verification tasks. As illustrated in Fig. 7, each 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.

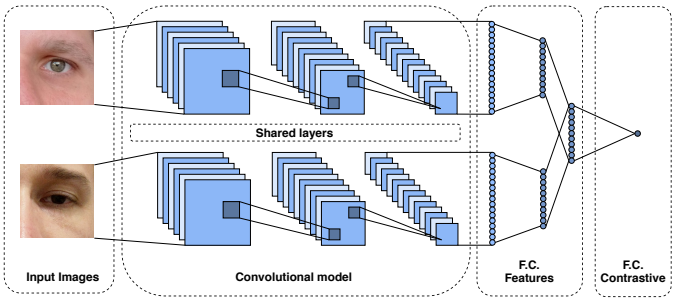


Fig. 7. Siamese CNN architecture. This model is composed of two twin branches of convolutional layers sharing their trainable parameters. The output computes a distance between the input image pairs.

As detailed in Table IV, we employ MobileNetV2 as the base model for each branch of the Siamese network. We use the contrastive loss [81], [82] in the training stage to compute the similarity between the input pair images.

TABLE IV
SIAMESE NETWORK ARCHITECTURE DESCRIPTION.

#	Layer	Connected to	Input	Output
0	branch_a (MobileNetV2 (88 layers))	–	$224 \times 224 \times 3$	256
1	branch_b (MobileNetV2 (88 layers))	–	$224 \times 224 \times 3$	256
2	dense	#0 and #1	512	256
3	Euclidean dist. / Contrastive loss	#2	256	1

As described in [82], let D_W be the Euclidean distance between two input vectors, the contrastive loss can be written as follows:

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

where

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

and P is the number of training pairs, $(Y,\vec{X_1},\vec{X_2})^i$ corresponds to the i-th label (Y) of the sample pair $\vec{X_1},\vec{X_2}$, 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 [81], [82] and has been employed for periocular recognition [83], [84] and iris recognition [57].

V. 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.

A. Experimental Setup

Inspired by several recent works [5], [28], [30], [31], [33], [54], [57], [68], [85], 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 [28], [31], [32], [35], 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 $224 \times 224 \times 3$, except for the InceptionResNet and Xception models, which have an input size of $299 \times 299 \times 3$. Note that the input dimensions are different because we are using pre-trained models and our fine-tuning process should respect the input size of the original architectures.

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 [4], [5], [28], [31], 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:

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

where A and B stand for the feature vectors.

Regarding the models explicitly developed for the verification tasks, i.e., the Siamese network and the Pairwise Filters network, 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 as diverse samples 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 5 repetitions for each fold, except for the Siamese and Pairwise filter networks, in which we ran only 3 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 Titan V GPU. All CNN models were implemented in python using the Tensorflow[3] and Keras[4] frameworks.

B. Benchmark results

This section presents the results obtained by each approach in the closed-world and open-world protocols. We also perform an ablation study on the Multi-task learning network to evaluate each task's influence in the identification mode. First, we show in Table V the size and the number of trainable parameters of each CNN model used as a benchmark. This information is from the models that we used on the closed-world protocol since they have more neurons on the last layer than the open-world protocol models.

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

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

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.

1) Closed-world protocol: In the closed-world protocol, we perform the benchmark for both the identification and verification tasks. All results are presented in Table VI. As can be seen, although MobileNetV2 is the smallest model in

[3]https://www.tensorflow.org/
[4]https://keras.io/

terms of size and trainable parameters, it achieved the best results for both identification and verification tasks. Hence, we used MobileNetV2 as the base model for the Multi-task, Siamese, and Pairwise Filters networks.

In general, the Multi-task model achieved the best results in terms of Rank 1, Rank 5, AUC, and EER. We highlight that we only explored the other tasks – age, gender, eye side, and mobile device model – at the training stage of this model. For the evaluation, we extracted the representations for the classification task and used it for the identification (using the softmax layer) and verification (using the cosine distance) tasks. The Siamese network obtained the worst results in the benchmark, while the Pairwise Filters network reached the higher Decidability index, indicating that it was the best at separating genuine and impostors distributions. However, it did not achieve the best results in terms of AUC and EER.

As stated in some previous works [28], [85], the models pre-trained for face recognition generally achieve best results than those pre-trained on the ImageNet dataset.

2) 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 VII.

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.

3) Multi-task Learning: The Multi-task model achieved the best results both 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 the closed-world protocol to evaluate the performance of both identification and verification. We also evaluated the results achieved by all models in each task.

According to Table VIII, 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. The gender and eye side classification tasks were handled well by all models, while the device 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, which can generate a bias during the training stage.

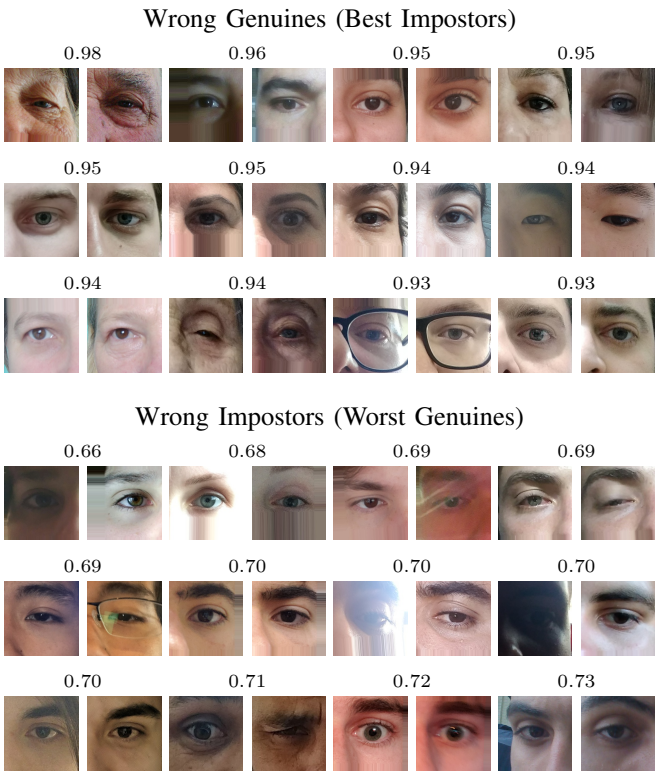


Fig. 8. Pairwise images wrongly classified by the model that obtained the best result in the verification task in the open-world protocol. Higher scores mean that the pair of periocular images is more likely to be genuine.

Note that in both closed-world and open-world protocols, we only explored the class prediction for the matching. However, as shown in Table VIII, 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.

4) 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. 8.

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 [4] 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,

TABLE VI
BENCHMARK RESULTS IN THE CLOSED-WORLD PROTOCOL FOR THE IDENTIFICATION AND VERIFICATION TASKS.

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	$\mathbf{84.32 \pm 0.71}$	$\mathbf{94.55 \pm 0.58}$	$\mathbf{99.96 \pm 0.01}$	$\mathbf{0.81 \pm 0.06}$	5.1978 ± 0.0340
Siamese	–	–	98.94 ± 0.22	4.86 ± 0.44	3.0005 ± 0.1871
Pairwise	–	–	99.44 ± 0.66	3.06 ± 1.84	$\mathbf{6.4503 \pm 1.2270}$

TABLE VII
BENCHMARK RESULTS IN THE OPEN-WORLD PROTOCOL FOR THE VERIFICATION TASK.

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**	$\mathbf{99.67 \pm 0.08}$	$\mathbf{2.81 \pm 0.39}$	3.9263 ± 0.0921
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	$\mathbf{4.4404 \pm 0.5834}$
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 VIII
RESULTS (%) FROM SEVERAL MULTI-TASK MODELS TRAINED TO PREDICT DIFFERENT TASKS.

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	$\mathbf{99.99 \pm 0.01}$
Multi-task (no age)	81.93 ± 0.99	93.51 ± 0.69	87.20 ± 0.63	–	97.65 ± 0.20	$\mathbf{99.99 \pm 0.01}$
Multi-task (no gender)	82.48 ± 0.64	93.55 ± 0.52	86.71 ± 0.54	83.17 ± 0.54	–	$\mathbf{99.99 \pm 0.01}$
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	$\mathbf{84.32 \pm 0.71}$	$\mathbf{94.55 \pm 0.58}$	$\mathbf{87.42 \pm 0.65}$	$\mathbf{84.34 \pm 0.71}$	$\mathbf{97.80 \pm 0.21}$	99.98 ± 0.02

mainly to the high intra-class variations.

VI. 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 ocular 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 had the most significant influence on the results. We stated that the mobile device model identification task was the most important one, followed by age range, gender, and eye side classification. 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 ocular 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 the 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. Therefore, there is still room for improvement in both identification and verification tasks.

ACKNOWLEDGMENT

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 Titan V GPU used for this research.

REFERENCES

[1] M. De Marsico, M. Nappi, and H. Proença, "Results from MICHE II - Mobile Iris CHallenge Evaluation II," *Pattern Recognition Letters*, vol. 91, pp. 3–10, May 2017.

[2] H. Proença and J. C. Neves, "IRINA: Iris recognition (even) in inaccurately segmented data," in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, Jul 2017, pp. 6747–6756.

[3] H. Proença and J. C. Neves, "A reminiscence of "mastermind": Iris/periocular biometrics by "in-set" CNN iterative analysis," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 7, pp. 1702–1712, July 2019.

[4] L. A. Zanlorensi, H. Proença, and D. Menotti, "Unconstrained periocular recognition: Using generative deep learning frameworks for attribute normalization," *arXiv preprint*, vol. arXiv:2002.03985, pp. 1–5, 2020.

[5] L. A. Zanlorensi, D. R. Lucio, A. S. Britto Jr., H. Proença, and D. Menotti, "Deep representations for cross-spectral ocular biometrics," *IET Biometrics*, vol. 9, pp. 68–77, 2020.

[6] K. B. Raja, R. Raghavendra, V. K. Vemuri, and C. Busch, "Smartphone based visible iris recognition using deep sparse filtering," *Pattern Recognition Letters*, vol. 57, pp. 33–42, May 2015.

[7] G. Santos, E. Grancho, M. V. Bernardo, and P. T. Fiadeiro, "Fusing iris and periocular information for cross-sensor recognition," *Pattern Recognition Letters*, vol. 57, pp. 52–59, May 2015.

[8] F. M. Algashaam, K. Nguyen, M. Alkanhal, V. Chandran, W. Boles, and J. Banks, "Multispectral periocular classification with multimodal compact multi-linear pooling," *IEEE Access*, vol. 5, pp. 14 572–14 578, 2017.

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

[10] M. Dobeš, L. Machala, P. Tichavský, and J. Pospíšil, "Human eye iris recognition using the mutual information," *Optik - International Journal for Light and Electron Optics*, vol. 115, no. 9, pp. 399–404, Jan 2004.

[11] M. S. Hosseini, B. N. Araabi, and H. Soltanian-Zadeh, "Pigment melanin: Pattern for iris recognition," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 4, pp. 792–804, 2010.

[12] M. De Marsico, M. Nappi, D. Riccio, and H. Wechsler, "Mobile Iris Challenge Evaluation (MICHE)-I, biometric iris dataset and protocols," *Pattern Recognition Letters*, vol. 57, pp. 17–23, 2015.

[13] A. Sequeira *et al.*, "Cross-Eyed - Cross-Spectral Iris/Periocular Recognition Database and Competition," in *International Conference of the Biometrics Special Interest Group*, vol. 260, Sept 2016, pp. 1–5.

[14] A. F. Sequeira *et al.*, "Cross-Eyed 2017: Cross-spectral iris/periocular recognition competition," in *IEEE International Joint Conference on Biometrics*, Oct 2017, pp. 725–732.

[15] P. R. Nalla and A. Kumar, "Toward more accurate iris recognition using cross-spectral matching," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 208–221, Jan 2017.

[16] H. Proença and L. A. Alexandre, "UBIRIS: A noisy iris image database," in *Image Analysis and Processing (ICIAP)*, 2005, pp. 970–977.

[17] H. Proença, S. Filipe, R. Santos, J. Oliveira, and L. A. Alexandre, "The UBIRIS.v2: A database of visible wavelength iris images captured on-the-move and at-a-distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1529–1535, aug 2010.

[18] C. N. Padole and H. Proença, "Periocular recognition: Analysis of performance degradation factors," in *IAPR International Conference on Biometrics (ICB)*, Mar 2012, pp. 439–445.

[19] A. Rattani, R. Derakhshani, S. K. Saripalle, and V. Gottemukkula, "ICIP 2016 competition on mobile ocular biometric recognition," in *IEEE International Conference on Image Processing – Challenge Session on Mobile Ocular Biometric Recognition*, Sept 2016, pp. 320–324.

[20] D. Menotti, G. Chiachia, A. Pinto, W. R. Schwartz, H. Pedrini, A. X. Falcão, and A. Rocha, "Deep representations for iris, face, and fingerprint spoofing detection," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 864–879, 2015.

[21] L. He, H. Li, F. Liu, N. Liu, Z. Sun, and Z. He, "Multi-patch convolution neural network for iris liveness detection," in *IEEE International Conf. on Biometrics Theory, Applications and Systems*, Sept 2016, pp. 1–7.

[22] P. Silva, E. Luz, R. Baeta, H. Pedrini, A. X. Falcao, and D. Menotti, "An approach to iris contact lens detection based on deep image representations," in *Conference on Graphics, Patterns and Images (SIBGRAPI)*, Aug 2015, pp. 157–164.

[23] D. R. Lucio, R. Laroca, L. A. Zanlorensi, G. Moreira, and D. Menotti, "Simultaneous iris and periocular region detection using coarse annotations," in *Conf. on Graphics, Patterns and Images*, 2019, pp. 178–185.

[24] E. Severo, R. Laroca, C. S. Bezerra, L. A. Zanlorensi, D. Weingaertner, G. Moreira, and D. Menotti, "A benchmark for iris location and a deep learning detector evaluation," in *International Joint Conference on Neural Networks (IJCNN)*, July 2018, pp. 1–7.

[25] D. R. Lucio, R. Laroca, E. Severo, A. S. Britto Jr., and D. Menotti, "Fully convolutional networks and generative adversarial networks applied to sclera segmentation," in *IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Oct 2018, pp. 1–7.

[26] C. S. Bezerra, R. Laroca, D. R. Lucio, E. Severo, L. F. Oliveira, A. S. Britto Jr., and D. Menotti, "Robust iris segmentation based on fully convolutional networks and generative adversarial networks," in *Conference on Graphics, Patterns and Images*, Oct 2018, pp. 281–288.

[27] Y. Du, T. Bourlai, and J. Dawson, "Automated classification of mislabeled near-infrared left and right iris images using convolutional neural networks," in *BTAS*, Sept 2016, pp. 1–6.

[28] E. Luz, G. Moreira, L. A. Zanlorensi Junior, and D. Menotti, "Deep periocular representation aiming video surveillance," *Pattern Recognition Letters*, vol. 114, pp. 2–12, 2018.

[29] T. Zhao, Y. Liu, G. Huo, and X. Zhu, "A deep learning iris recognition method based on capsule network architecture," *IEEE Access*, vol. 7, pp. 49 691–49 701, 2019.

[30] K. H. Diaz, F. Alonso-Fernandez, and J. Bigun, "Spectrum translation for cross-spectral ocular matching," *arXiv arXiv:2002.06228*, 2020.

[31] L. A. Zanlorensi, E. Luz, R. Laroca, A. S. Britto Jr., L. S. Oliveira, and D. Menotti, "The impact of preprocessing on deep representations for iris recognition on unconstrained environments," in *Conference on Graphics, Patterns and Images (SIBGRAPI)*, Oct 2018, pp. 289–296.

[32] P. H. Silva, E. Luz, L. A. Zanlorensi, D. Menotti, and G. Moreira, "Multimodal feature level fusion based on particle swarm optimization with deep transfer learning," in *2018 Congress on Evolutionary Computation (CEC)*, July 2018, pp. 1–8.

[33] K. Hernandez-Diaz, F. Alonso-Fernandez, and J. Bigun, "Cross-spectral periocular recognition with conditional adversarial networks," 2020.

[34] H. Proença and L. A. Alexandre, "Toward covert iris biometric recognition: Experimental results from the NICE contests," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 798–808, 2012.

[35] L. A. Zanlorensi, R. Laroca, E. Luz, A. S. Britto Jr., L. S. Oliveira, and D. Menotti, "Ocular recognition databases and competitions: A survey," *arXiv preprint*, vol. arXiv:1911.09646, pp. 1–20, 2019.

[36] P. J. Phillips, K. W. Bowyer, P. J. Flynn, X. Liu, and W. T. Scruggs, "The iris challenge evaluation 2005," in *IEEE International Conference on Biometrics: Theory, Applications and Systems*, Sep. 2008, pp. 1–8.

[37] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe, "FRVT 2006 and ICE 2006 large-scale experimental results," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 831–846, 2010.

[38] S. Shah and A. Ross, "Generating synthetic irises by feature agglomeration," in *International Conf. on Image Processing*, 2006, pp. 317–320.

[39] J. Zuo, N. A. Schmid, and X. Chen, "On generation and analysis of synthetic iris images," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 1, pp. 77–90, 2007.

[40] V. Ruiz-Albacete, P. Tome-Gonzalez, F. Alonso-Fernandez, J. Galbally, and J. Ortega-Garcia, "Direct attacks using fake images in iris verification," in *Biometrics and Identity Management*, 2008, pp. 181–190.

[41] A. Czajka, "Database of iris printouts and its application: Development of liveness detection method for iris recognition," in *International Conf. on Methods Models in Automation Robotics*, Aug 2013, pp. 28–33.

[42] P. Gupta, S. Behera, M. Vatsa, and R. Singh, "On iris spoofing using print attack," in *International Conference on Pattern Recognition (ICPR)*, Aug 2014, pp. 1681–1686.

[43] N. Kohli, D. Yadav, M. Vatsa, R. Singh, and A. Noore, "Detecting medley of iris spoofing attacks using DESIST," in *IEEE Intl. Conf. on Biometrics Theory, Applications and Systems*, Sep. 2016, pp. 1–6.

[44] S. E. Baker, A. Hentz, K. W. Bowyer, and P. J. Flynn, "Degradation of iris recognition performance due to non-cosmetic prescription contact lenses," *Computer Vision and Image Understanding*, vol. 114, no. 9, pp. 1030–1044, Sept 2010.

[45] N. Kohli, D. Yadav, M. Vatsa, and R. Singh, "Revisiting iris recognition with color cosmetic contact lenses," in *International Conference on Biometrics (ICB)*, vol. 1, Jun 2013, pp. 1–7.

[46] J. S. Doyle, K. W. Bowyer, and P. J. Flynn, "Variation in accuracy of textured contact lens detection based on sensor and lens pattern," in *BTAS*, Sept 2013, pp. 1–7.

[47] J. S. Doyle and K. W. Bowyer, "Robust detection of textured contact lenses in iris recognition using BSIF," *IEEE Access*, vol. 3, pp. 1672–1683, 2015.

[48] S. P. Fenker and K. W. Bowyer, "Analysis of template aging in iris biometrics," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Jun 2012, pp. 45–51.

[49] S. E. Baker, K. W. Bowyer, P. J. Flynn, and P. J. Phillips, *Template Aging in Iris Biometrics*. Springer London, 2013, ch. 11, pp. 205–218.

[50] K. W. Bowyer, K. Hollingsworth, and P. J. Flynn, "Image understanding for iris biometrics: A survey," *Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 281–307, 2008.

[51] H. Proença and J. C. Neves, "Segmentation-less and non-holistic deep-learning frameworks for iris recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 1–10.

[52] J. Daugman, "High confidence visual recognition of persons by a test of statistical independence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 11, pp. 1148–1161, 1993.

[53] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[54] N. Reddy, A. Rattani, and R. Derakhshani, "Comparison of deep learning models for biometric-based mobile user authentication," in *IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2018, pp. 1–6.

[55] A. K. Jain and A. Ross, *Introduction to Biometrics*. Springer US, 2008, pp. 1–22.

[56] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[57] K. Wang and A. Kumar, "Cross-spectral iris recognition using cnn and supervised discrete hashing," *Pattern Recognition*, vol. 86, pp. 85–98, 2019.

[58] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin dags for multiclass classification," in *International Conference on Neural Information Processing Systems (NIPS)*, Nov 1999.

[59] T. Hastie, S. Rosset, J. Zhu, and H. Zou, "Multi-class adaboost," *Statistics and Its Interface*, vol. 2, no. 3, pp. 349–360, 2009.

[60] G. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 2, pp. 513–529, 2012.

[61] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, May 2015.

[62] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference (BMVC)*, 2015, pp. 1–12.

[63] T. Zhao, Y. Liu, G. Huo, and X. Zhu, "A deep learning iris recognition method based on capsule network architecture," *IEEE Access*, vol. 7, pp. 49 691–49 701, 2019.

[64] S. S. Behera, S. S. Mishra, B. Mandal, and N. B. Puhan, "Variance-guided attention-based twin deep network for cross-spectral periocular recognition," *Image and Vision Computing*, p. 104016, 2020.

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

[66] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," *CoRR*, 2017.

[67] A. Boyd, A. Czajka, and K. Bowyer, "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)*, 2019, pp. 1–9.

[68] F. Boutros, N. Damer, K. Raja, R. Ramachandra, F. Kirchbuchner, and A. Kuijper, "Fusing iris and periocular region for user verification in head mounted displays," in *IEEE International Conference on Information Fusion (FUSION)*, 2020, pp. 1–8.

[69] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conf. on Computer Vision*, 2016, pp. 630–645.

[70] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *ICLR 2016 Workshop*, 2016.

[71] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[72] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[73] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[74] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[75] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[76] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[77] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, p. 41–75, Jul. 1997.

[78] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[79] N. Liu, M. Zhang, H. Li, Z. Sun, and T. Tan, "Deepiris: Learning pairwise filter bank for heterogeneous iris verification," *Pattern Recognition Letters*, vol. 82, pp. 154 – 161, 2016.

[80] J. Bromley, I. Guyon, Y. LeCun, E. .Säckinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," in *Intl. Conf. on Neural Information Processing Systems*, 1993, p. 737–744.

[81] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 539–546.

[82] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2006, pp. 1735–1742.

[83] Z. Zhao and A. Kumar, "Improving periocular recognition by explicit attention to critical regions in deep neural network," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 12, pp. 2937–2952, 2018.

[84] S. S. Behera, B. Mandal, and N. B. Puhan, "Twin deep convolutional neural network-based cross-spectral periocular recognition," in *2020 National Conference on Communications (NCC)*, 2020, pp. 1–6.

[85] A. Boyd, A. Czajka, and K. Bowyer, "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)*, 2019, pp. 1–9.