Deep-PRWIS: Periocular Recognition Without the Iris and Sclera Using Deep Learning Frameworks

Hugo Proença, *Senior Member, IEEE* and João C. Neves, *Member, IEEE*

***Abstract*—This work is based on a disruptive hypothesis for periocular biometrics: in visible-light data, the recognition performance is optimized when the components inside the ocular globe (the iris and the sclera) are simply discarded, and the recogniser's response is exclusively based in information from the surroundings of the eye. As major novelty, we describe a processing chain based on convolution neural networks (CNNs) that defines the regions-of-interest in the input data that should be privileged in an implicit way, i.e., without masking out any areas in the learning/test samples. By using an ocular segmentation algorithm exclusively in the learning data, we separate the ocular from the periocular parts. Then, we produce a large set of "multi-class" artificial samples, by interchanging the periocular and ocular parts from different subjects. These samples are used for data augmentation purposes and feed the learning phase of the CNN, always considering as label the ID of the periocular part. This way, for every periocular region, the CNN receives multiple samples of different ocular classes, forcing it to conclude that such regions should not be considered in its response. During the test phase, samples are provided without any segmentation mask and the network *naturally* disregards the ocular components, which contributes for improvements in performance. Our experiments were carried out in full versions of two widely known data sets (UBIRIS.v2 and FRGC) and show that the proposed method consistently advances the state-of-the-art performance in the *closed-world* setting, reducing the EERs in about 82% (UBIRIS.v2) and 85% (FRGC) and improving the Rank-1 over 41% (UBIRIS.v2) and 12% (FRGC).**

***Index Terms*—Soft Biometrics, Visual Surveillance, Homeland Security.**

CONVOLUTIONAL neural networks (CNNs) have become extremely popular in many computer vision tasks, from image segmentation [10], to detection [23] and classification [9]. The property of shift/space invariance gives them the biological inspiration and simultaneously keeps the number of weights relatively small, making learning a feasible task. Being data-driven models, CNNs do not depend on human efforts to specify the image features, upon the availability of large amounts of learning data.

I. Introduction

In the biometrics domain, the covert recognition of humans (outdoor and non-cooperative) remains to be achieved, and will be a breakthrough in security/forensics applications. Here, the periocular region is a trade-off between using the iris and the face, with encouraging performance levels reported in the literature. However, as it is illustrated in Fig. 1, it should be considered that:

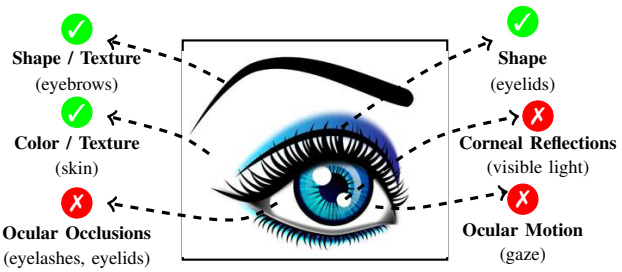

Fig. 1. Schema of the components in the ocular/periocular regions, with the three major factors that reduce the reliability of the ocular components for biometric recognition in covert mode: 1) eye gaze; 2) iris/sclera occlusions; and 3) corneal reflections.

- when imaged under visible-light, the iris (particularly) and the sclera are prune to corneal reflections, resulting in the so-called Purkinje images;
- along with the body and head movements, the components in the ocular globe are subjected to an additional motion source (eye gaze) that increases the probabilities of acquiring blurred data;
- the iris and the sclera are often partially occluded, due to eyelids and eyelashes movements;

According to the above points, this paper describes a periocular recognition algorithm to work in poor quality visible-light data, relying on CNNs to model complex data patterns. The key is a data augmentation strategy based in multi-class regions swapping, that implicitly induces the CNN to consider that some regions in the input data are not reliable for classification purposes. This is seen as a novel way to provide prior knowledge to this kind of networks, considerably improving performance without requiring extra amounts of learning data. Note that this strategy can be easily generalized to other object classification problems, i.e., to any case where the discriminability provided by the different image components varies substantially and there is not enough learning data available to expect the autonomous inference of such conclusion by the network.

The workflow is illustrated in Fig. 2 (*Learning* box): by using an ocular segmentation algorithm [17], we create a binary mask $\mathbf{B}$ that discriminates between the ocular $\mathbf{O}$ (iris and sclera) and the remaining components $\mathbf{P}$ (henceforth designated as periocular, including the eyebrows, eyelids, eyelashes and skin) in each learning sample. Next, a set of artificial samples is created, interchanging the ocular and periocular parts from

Authors are with the IT: Instituto de Telecomunicações, Department of Computer Science, University of Beira Interior, Covilhã, Portugal, E-mail: {hugomcp, jcneves}@di.ubi.pt. This work was supported by FCT project UID/EEA/50008/2013.

Manuscript received ? ?, 2017; revised ? ?, ?.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2017.2771230, IEEE Transactions on Information Forensics and Security

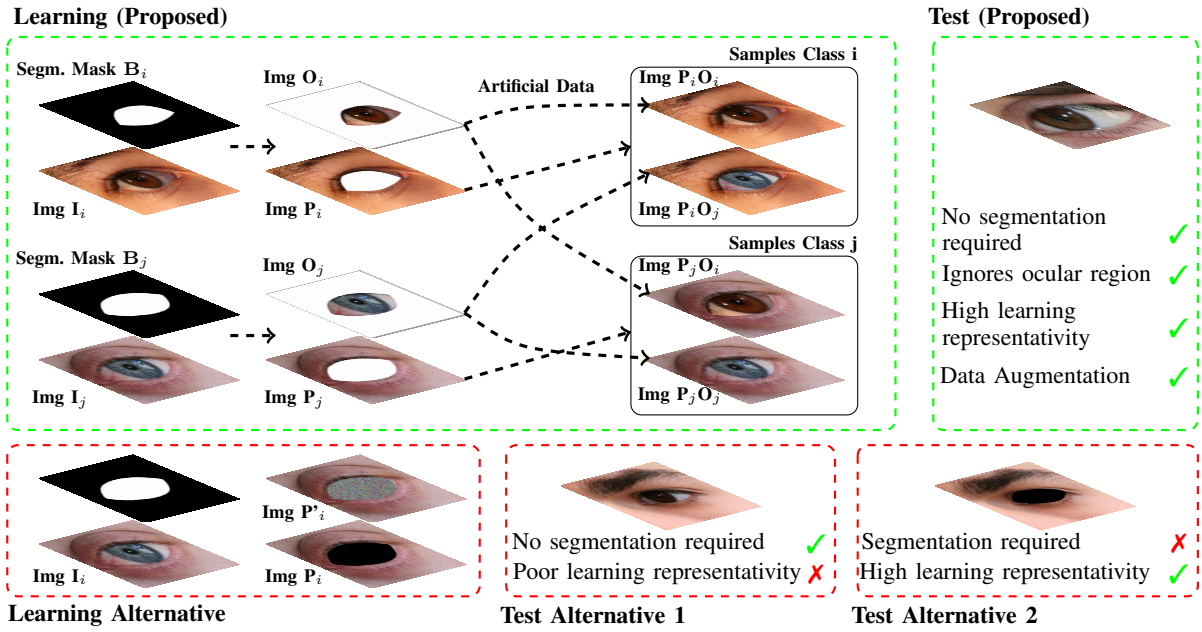

Fig. 2. Schema of the strategy used to implicitly force the CNN to disregard regions in the input data. Creating artificial "multi-class" samples that keep as label the ID of the periocular part, leads the network to consider that ocular patterns are meaningless for biometric recognition. This yields four properties (given at the top-right corner), which will not be verified for any other combination of learning/testing strategies (given at the bottom part of the figure).

different subjects, but always considering as label the ID provided by the periocular part. This way, during the learning phase, the CNN receives, for each periocular part, samples of different ocular classes, forcing it to conclude that such regions should not be considered in its response (i.e., the ID). During the test phase (*Test* box), samples are provided to the network without any segmentation mask, yielding four key properties: 1) the CNN testing performance is not conditioned by the effectiveness of the segmentation step, known to be a primary error source in computer vision tasks; 2) the CNN naturally ignores the ocular components, focusing in the most discriminating information; 3) the learning and test data have similar appearance, which contributes to the CNN's generalization capability; and 4) from a data augmentation perspective, the set of artificial samples provided to the network also improves the CNN performance. As shown in the bottom part of Fig. 2, any other combination of learning/test data (using explicit region masking) will not keep these four properties simultaneously.

As outcome of this work, the resulting periocular recogniser outperforms consistently the state-of-the-art, decreasing the EERs and improving the Rank-1 values with respect to the baseline methods. Note that these results were obtained in two widely known data sets and using the entire set of images in both sets, i.e., without disregarding even the poorest quality samples.

The remainder of this paper is organised as follows: Section II summarises the periocular biometrics research, and Section III describes our method. In Section IV we discuss the obtained results and the conclusions are given in Section V.

II. Related Work

The pioneer work on periocular biometrics was due to Park *et al*. [14] (extended in [15]). They consider the iris as the reference for defining the ROI, described by HoG, LBP and SIFT descriptors. The ℓ_2 norm is the distance measure for each descriptor, with results fused at the score level, by linear combination. This work provided the basis for a large number of subsequent methods: Mahalingam and Ricanek Jr. [11] apply multi-scale, patch-based LBP descriptors, using iris center for data alignment. Ross *et al*. [21] use HoGs to extract the global image information, SIFT to extract local edge anomalies, and probabilistic deformation models to handle non-linear deformations, with the sum rule combining the dissimilarity scores. Bharadwaj *et al*. [2] apply global descriptors (GIST and circular LBPs), each one compared using the Chi-square distance. Scores are also linearly combined. Woordard *et al*. [26] fuse local appearance-based feature descriptors to 2D color histograms (red and green channels), compared using the city-block (LBP) and Bhattacharya (color histograms) distances. Joshi *et al*. [8] describe the periocular information by mean of a bank of complex Gabor filters, while Tan and Kumar [24] evaluate the effectiveness of SIFT, GIST, LBP, HoG and Leung-Malik Filters texture descriptors to provide discriminating information on periocular data. The singularity of Nie *et al*. [12]'s work is to combine this kind of classical approach to a convolutional restricted Boltzmann machine, which enables to obtain the probability distributions in the periocular data, discriminated by metric learning and SVMs.

Additional approaches are due to Chen and Ferryman [5], which fuse 2D to 3D data, masking out the ocular region from the encoding and comparison process. Raghavendra *et al*. [20]

exploit the light-field data acquisition technology to produce sharp images for iris and periocular recognizers, with scores linearly combined. Aiming at cross-spectral recognition, Cao and Schmid [4] convolve the periocular region with a bank of Gabor filters, from which phase and magnitude components are described by HoGs and histograms of LBPs descriptors. Features are concatenated and compared using the I-divergence measure. As an anti-counterfeit measure, Proença [19] propose an ensemble made of two disparate experts: one analysing the iris texture and the other one parameterizing the shape of eyelids and analysing the surrounding skin. Both experts provide independent responses and do not share particularly sensitivity to any image covariate.

In terms of deep learning-based approaches, Zhao and Kumar [27] use a CNN for periocular recognition (as we do). The novelty is to consider explicit semantic information to extract more comprehensive periocular features, helping the CNN to improve performance. Refer to the surveys on periocular biometrics due to Alonso-Fernandez and Bigun [1] and Nigam *et al*. [13], for additional information about the periocular biometrics research.

Recently, particular attention has been paid to the recognition of cross-spectral iris/periocular data, i.e., when the pairs of images to be compared were acquired using different light wavelengths (typically near infra-red and visible). Several approaches were published in this field, with some results and relevant methods described in [22].

III. Proposed Method

A. Deep Learning Architecture

We use one of the most popular deep learning architectures for image classification: Convolution Neural Networks (CNNs), which are a biologically inspired variant of multilayer perceptron networks (MLPs) particularly suitable for image classification. By making some assumptions about the nature of the input data (e.g., stationarity of statistics and locality of pixel dependencies), CNNs have much fewer connections than MLPs, making learning a feasible task. In particular, we adopt a CNN architecture based in AlexNet [9], shown in Fig. 3. This classical architecture boosted the popularity of deep learning frameworks for image classification, and is know to constitute a good trade-off between the number of model parameters and the generalisation capabilities of the final solution. The idea is to start by extracting features of increasing complexity at the deeper layers of the network (using convolution layers), which feed the final fully connected layers that provide the final response. At the same time, max pooling and dropout layers keep the number of parameters relatively low while not compromising the generalization capabilities of the network. Our input data are $150 \times 200 \times 3$ RGB images that pass through convolution (at first), max-pooling, dropout and fully connected layers. All the convolutional layers are adjacent to Rectified Linear Unit (ReLU) activation functions, being the ith output channel $\mathbf{y}^{(i)}$ given by:

$$\mathbf{y}^{(i)} = \max\Big(\sum_{j=1}^{k} \mathbf{b}^{(ij)} + \mathbf{w}^{(ij)} * \mathbf{x}^{(j)}, \mathbf{0}_p\Big), \qquad (1)$$

where $\max(.,\mathbf{0}_p)$ is the component-wise maximum operator, $\mathbf{0}_p = [0,\ldots,0]^T$ is an $p \times 1$ vector with all elements equal to zero, $\mathbf{b}^{()}$ and $\mathbf{w}^{()}$ are the bias and weight terms tuned during the learning phase and $\mathbf{x}$ represents the layer inputs. The max-pooling layers operate independently in each depth slice of the input and take the maximum value over square patches. Finally, dropout layers set to zero the output of each neuron during the learning step with probability r, avoiding that they contribute to the forward pass and participate in back-propagation.

In our model, the first convolutional layer has 128 kernels (5×5), using stride and padding of two pixels. Next, a max-pooling and a dropout layer feed the second and third convolutional layers composed of 256 kernels (5×5, two pixels of stride and padding). Again, a max-pooling shrinks the volume data and then two convolutional layers with output size equal to the input are applied (256 kernels of size 3×3, stride and padding equal to one). Before the fully connected layers, data pass through a convolution layer (with 512 kernels of size 3×3, stride and padding equal to one) a max pooling and a dropout layer, yielding $9 \times 12 \times 512 = 55,296$ features entering in the fully connected layers. Another dropout layer is used before the soft-max layer, that produces a vector of c positive elements corresponding to the probability for each class label:

$$P(y = j|\mathbf{x}) = \frac{e^{\mathbf{x}^T \mathbf{w}_j}}{\sum_k e^{\mathbf{x}^T \mathbf{w}_k}}. \qquad (2)$$

According to the output of the soft-max layer, the label prediction is the class with the highest probability among the c possibilities: $\hat{y} = \arg\max_j P(y = j|\mathbf{x})$. The CNNs were trained using the stochastic gradient descent (SGD) algorithm, with a batch size of 256 samples. As preprocessing step, the mean of the learning data was subtracted from all samples. The learning rate was $1e^{-3}$, with a momentum of 0.9 and a weight decay of $5e^{-4}$. The number of iterations in each experiment was set to 100. All weights in the CNN were initialised according to Glorot and Bengio's [6] method.

B. Data Augmentation

1) Ocular/Periocular Regions Swapping: Let $\mathbf{I}_i$ and $\mathbf{I}_j$ be $150 \times 200 \times 3$ RGB images from two different subjects. Using the segmentation method described in [17], we obtain two binary masks $\mathbf{B}_i$ and $\mathbf{B}_j$ (150×200 pixels) that discriminate between the ocular (iris and sclera) and the periocular components in $\mathbf{I}_.$. Let $\mathbf{O}_.$ and $\mathbf{P}_.$ denote the ocular and periocular parts of $\mathbf{I}_.$. The goal is to create an artificial sample $\mathbf{P}_i\mathbf{O}_j$ composed of the periocular region of $\mathbf{I}_i$ overlapping the ocular part of $\mathbf{I}_j$, which requires to find the scale and translation parameters, such that $\mathbf{O}_j$ optimally fits the ocular whole of $\mathbf{P}_i$. Let $\mathbf{b}_.$ be the $n \times 1$ vectorized version of $\mathbf{B}_.$ (n=30,000). The convolution "*" between $\mathbf{b}_i$ and $\mathbf{b}_j$ is given in matrix form by:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2017.2771230, IEEE Transactions on Information Forensics and Security

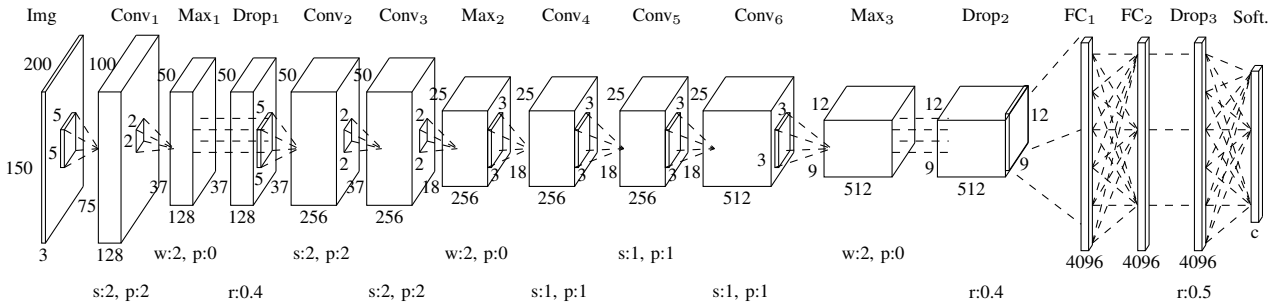

Fig. 3. Structure of the convolutional neural network used in image classification. Six convolutional layers, three max-pooling and dropout layers are used before the (three) fully connected and the soft-max layer, that estimates the sample identity. "s:_" denotes the stride, "p:_" specifies the padding, "w:_" is the square neighborhood used in max-pooling layers and "r:_" defines the dropout rate. Note that all convolution layers also include "ReLU" non-linear transfer functions.

$$\mathbf{b}_i * \mathbf{b}_j = \mathbf{T}(\mathbf{b}_i)\ \mathbf{b}_j, \tag{3}$$

being $\mathbf{T}(\mathbf{b}_i)$ the Toeplitz matrix of $\mathbf{b}_i$:

$$\mathbf{T}(\mathbf{b}_i) = \left.\begin{bmatrix} \mathbf{b}_i & 0 & \dots & 0 \\ 0 & \mathbf{b}_i & \dots & \vdots \\ \vdots & \vdots & \vdots & 0 \\ 0 & 0 & 0 & \mathbf{b}_i \end{bmatrix}\right\} (2n-1) \times n. \tag{4}$$

Let $\mathbf{1}_{2n-1} = [1, \dots, 1]^T$ be the $(2n-1) \times 1$ vector having all elements equal to one. According to this formulation, the value of:

$$\mathbf{1}_{2n-1}^T \big(\mathbf{T}(\mathbf{b}_i)\ \mathbf{b}_j\big), \tag{5}$$

directly corresponds to the agreement of the ocular parts of $\mathbf{B}_i$ and $\mathbf{B}_j$ (i.e., their white regions). Let $\neg\mathbf{b}_.$ be the negative version of $\mathbf{b}_.$. As we are interested in maximise the "ocular" $\leftrightarrow$ "ocular" and "periocular" $\leftrightarrow$ "periocular" positions agreement, while minimising the "ocular" $\leftrightarrow$ "periocular" disagreements between masks, the unknown scale and translation parameters $\boldsymbol{\alpha} = [\alpha_s, \alpha_x, \alpha_y]$ that optimally overlap $\mathbf{P}_i$ and $\mathbf{O}_j$ are found by:

$$\begin{aligned} \hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \mathbf{1}_{2n-1}^T \Big(\big(\mathbf{T}(\neg\mathbf{b}_i) - \mathbf{T}(\mathbf{b}_i)\big)\big(\mathbf{b}_j^{(\boldsymbol{\alpha})} - \neg\mathbf{b}_j^{(\boldsymbol{\alpha})}\big)\Big) \\ \text{s.t. } ||[\alpha_x, \alpha_y]||_\infty \leq \kappa_1 \ \wedge\ \frac{1}{\kappa_2} \leq \alpha_s \leq \kappa_2, \end{aligned} \tag{6}$$

where $\mathbf{b}_j^{(\boldsymbol{\alpha})}$ is the translated and scaled version of $\mathbf{b}_j$, and κ_i avoid anatomically bizarre solutions ($\kappa_1 = 50, \kappa_2 = 3$ in our experiments). According to this formulation, (6) is a constrained optimization problem with inequality constraints, solved as described in [3]. In practice, we find the displacement (α_x, α_y) of the scaled (by α_s) version of $\mathbf{O}_j$ that optimally fits $\mathbf{P}_i$, yielding artificial samples that are realistic and visually pleasant. Examples of this overlapping procedure are given in Fig. 4, with the leftmost column showing samples of the UBIRIS.v2 set, and the remaining columns displaying artificial samples composed of the periocular region at left and different ocular parts.

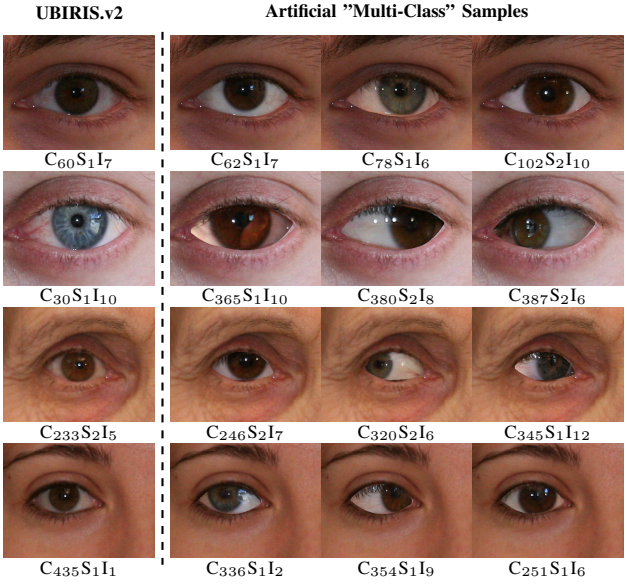

Fig. 4. Left column: UBIRIS.v2 samples. At right: Artificial "multi-class" samples composed of the periocular region given at left and the ocular parts given below each image. Note that the periocular region in these "multi-class"' samples in each row is the same.

2) Spatial and Color Transforms: Additionally, two other label-preserving transformations were used for data augmentation purposes. At first, to simulate the scale and translation samples inconsistency, patches of scale $[0.75, 0.90]$ (values drew uniformly) were randomly cropped from the learning set, as illustrated in the upper row of Fig. 5. Second, to get a color transform, we found the principal components of the RGB values in all pixels of the learning data and created new versions of the images by adding to each pixel multiples of the largest eigenvectors, with magnitude equal to the corresponding eigenvalues [9]:

$$\mathbf{x}^{(\text{new})} = \mathbf{x}^{(\text{old})} + [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3] \left(\alpha \odot [\lambda_1, \lambda_2, \lambda_3]^T\right), \tag{7}$$

with $\odot$ denoting the element-wise multiplication, $\mathbf{v}$ and λ denoting the eigenvectors and eigenvalues of the learning data covariance matrix and $\alpha \in \mathbf{R}^3$ being randomly drew from the Gaussian $\mathcal{N}(0, 0.1)$. Examples of the resulting images are given in the bottom row of Fig. 5.

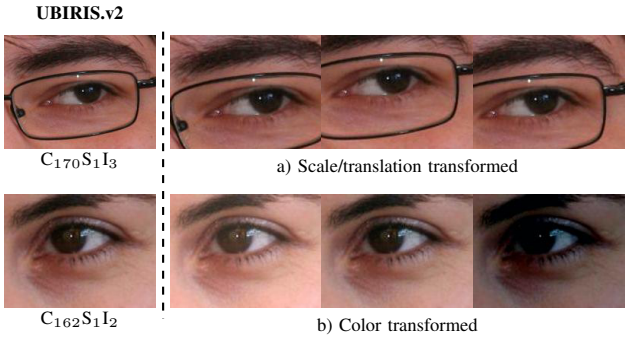

Fig. 5. Examples of the scale, translation and color transforms used. The upper row illustrates the randomly cropped patches, and the bottom row shows changes in color, obtained by adding multiples of the principal component vectors to each image pixel.

IV. Results and Discussion

A. Open *vs.* Closed-*World Settings*

When using classification models such as the ones in this paper (CNNs), an important decision to make is about the *working mode* most suitable for the model. In particular, it should be defined if the resulting system is expected to work in the *open-world* or *closed-world* mode, i.e., depending if the system possesses at learning time samples from all the classes that will be seen at runtime or not. In case of CNN-based classification tasks, the *closed-world* mode enables to use the output of the neurons in the final soft-max layer as the probabilities for each class label. In opposition, in the *open-world* mode, the number of different classes seen at runtime is not known, and the soft-max layer cannot be used. Instead, the output of the final convolution layer is typically used as feature descriptor and the ℓ_2 norm gives the distance between two feature sets, discriminating between genuine or impostor comparisons. In our experiments, having observed some preliminary results about the recognition performance of our method in both modes, we decided to focus exclusively in the *closed-world* scenario, i.e., assuming that the set of identities to be recognized is known in learning time. As an example, the latter setting corresponds to a watch-list identification problem, where the goal is to find subjects in a short list among a crowd.

B. Datasets and Experimental Protocol

Two datasets were selected for our experiments: 1) the UBIRIS.v2 [18], which is typically used for iris and periocular recognition experiments. All images of this set were used (11,102 images from 522 different eyes), regardless the extreme poor quality of some of them. Images have $150 \times 200 \times 3$ pixels and are represented in the RGB color space; and 2) the Face Recognition Grand Challenge [16] (FRGC) set, released by the National Institute of Standards and Technology (NIST). Again, all the 24,946 RGB samples in this set (with periocular regions cropped and resized into $150 \times 200 \times 3$ pixels) were considered. Cropping the left/right eye regions from each image yields a total of 894 classes. Examples of some of the poorest quality images used in our evaluation are given in Fig. 6.

All experiments were conducted according to a bootstrapping-like strategy, which is widely adopted in biometric recognition experiments (e.g. [7]). Having n images available, the bootstrap randomly selects (without replacement in our case) $0.9n$ images, creating a sample composed by 90% of the available data. This sample is disjointly divided into two subsets: 80% for learning purposes and the remaining 20% for performance evaluation. Note that we manually verified that the learning subsets, both for UBIRIS and FRGC, contain images from all the subjects (classes) in the data set, assuring that the *closed-world* assumption is satisfied.

The bootstrapping-like draw was repeated 10 times per data set, creating 10 subsets of each one. Next, the recognition experiments (model learning and performance evaluation) were carried out in each subset, which enabled us to obtain the average and standard deviation performance values at all operating points for both the UBIRIS.v2 and FRGC sets. These are the values that are reported in Table I and in all ROC and RANK-N plots (with the lines providing the average performance and the shadowed regions denoting the standard deviations at each position).

For all our experiments, the MATLAB® programming language was chosen, with the *MathConvNet* [25] toolbox used to implement our CNN models. Also, a NVIDIA® *Titan X* GPU was used to speed-up the learning processes, with 12GB memory and 3,072 CUDA cores.

Fig. 6. Datasets used in our empirical evaluation. The upper row regards the UBIRIS.v2 set, with five major degradation factors: iris occlusions, reflections, varying pose, glasses and poor lighting conditions. The bottom row regards the FRGC set, where the major degradation factors are image blur, poor resolution and bad lighting.

C. Data Augmentation: Performance Optimization

For performance optimisation, one important point is the amount of artificial data required with respect to the original number of images, avoiding the unrealistic "*as large as possible*" paradigm. Having three types of data augmentation strategies (scale/translation transforms, color transforms and regions swapping), the goal here is to perceive the amounts of data above which performance improvements are residual, if any. To get that threshold, we augmented the data from one to 64x (considering the original number of images per

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2017.2771230, IEEE Transactions on Information Forensics and Security

data set), and repeated the learning / performance evaluation steps. As given in the left part of Fig. 7, in the case of the UBIRIS.v2 set, performance consistently improves up to the point where the augmented data is about 32x the original samples (i.e., using approximately 350,000 artificial images), above where improvements in performance decrease and start to be residual. Regarding the FRGC set, the stabilization in performance was observed a slight earlier, i.e., when the amount of augmented data was 8x to 16x the number of original images (corresponding to approximately 400,000 artificial images).

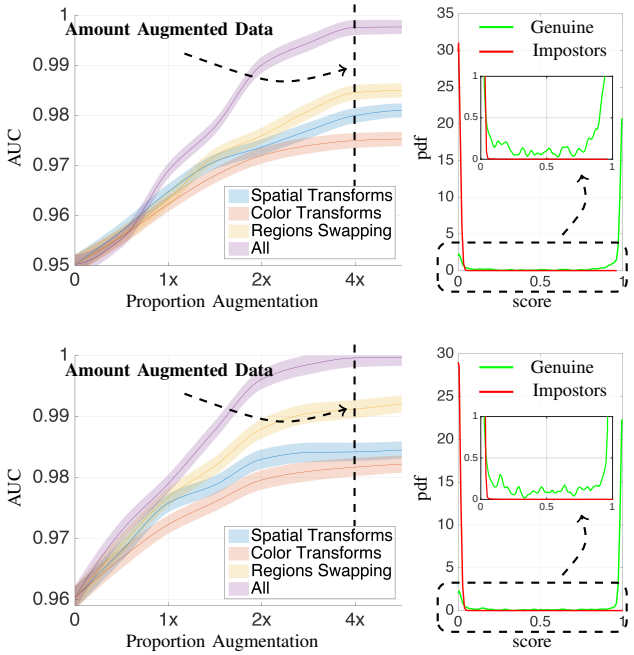

Fig. 7. Left plot: variations in recognition performance with respect to the amount of augmented data, with respect to the number of samples in the data set. Right plot: decision environment of the responses given by the neurons in the final layer of the CNN, distinguishing between the genuine (green) and impostor (red) class scores. The zoomed-in region turns particularly evident the recogniser bias. The upper row regards the UBIRIS.v2 data set, whereas the bottom row gives the corresponding values for the FRGC set.

In terms of the typical scores generated by the CNNs, the right side of Fig. 7 plots the genuine/impostor scores likelihood densities for the UBIRIS.v2 (upper row) and FRGC sets. The zoomed-in region turns particularly evident the classifier bias, in which errors are most times due to false negative responses, i.e., the genuine distribution has non-zero densities along the unit interval, which doesn't happen for the impostor scores, where non-residual densities appear exclusively near the zero value. In practice, this yields one important requirement for biometric systems to work in degraded data: the residual probability of observing false-matches. In these cases, regardless the system's sensitivity, it can be stated with full confidence that any reported match is genuine.

According to these results, in all subsequent experiments we kept the amount of augmented data as 32x the original data set and compared our algorithm's performance to three baseline strategies: the works due to Zhao and Kumar [27], Tan and Kumar [24] and Proença [19]. These techniques are summarized in Sec. II and were selected because they report the state-of-the-art performance ([27] and [24]), use techniques that are similar to ours ([27]) and were designed to work in similar conditions to our method ([19]). However, note that the Zhao and Kumar [27]'s method was designed to work in a more challenging scenario, corresponding to the *open-world* operating mode.

D. All *vs.* Periocular *vs.* Ocular *CNNs*

As stated above, the underlying hypothesis in this paper is that periocular recognition performance improves when the less reliable components (the iris and the sclera) are discarded by the CNN. Fig. 9 compares the performance attained when using all the image components (iris, sclera, eyelids, eyelashes, eyebrows and skin), and when the components inside the ocular globe are implicitly discarded (according to the data augmentation strategy described in Sec. III-B1). As a reference, we also show the performance obtained by the complementary configuration (i.e., using only the iris and the sclera), which is done simply by using the ID of the ocular part in each augmented sample. As can be seen both in the ROC and Rank-N curves, the best performance is attained when the ocular components are discarded, with solid differences in performance and non-overlapping confidence intervals. The small reliability of the iris and sclera for biometric recognition in visible-light environments is confirmed by the performance attained by the *Ocular* classifier, with performance levels dramatically poorer than the other two configurations (*All* and *Periocular*). Results in this Fig. regard exclusively the UBIRIS.v2 set, even though almost overlapping differences in performance were observed for FRGC. As these results are clearly redundant to those provided for UBIRIS.v2, we decided not to include them in the paper.

Moreover, the different features learned by the CNNs when using only some of the components are evident by analyzing the average magnitude of the 512 (9×12) filters tuned by the SGD algorithm immediately before the fully connected layers, i.e., the first point in the CNN where the filters coefficients have a bijective correspondence to input image positions. Results are given in Fig. 10 for three types of CNNs: in a) the CNN learns from all the regions of the input data, i.e., without using the image overlapping strategy described in Sec. III-B1; in b) only the ocular regions are considered by the CNN; and in c) only the periocular regions are considered. It can be seen that the average magnitude of the coefficients spreads evenly in a) and has obvious valleys in the regions that are implicitly demanded to be discarded, according to the data augmentation strategy used. This confirms that the CNNs are actually disregarding or, at least, giving less importance to the information in these regions.

E. State-of-the-Art Results Comparison

The ROC curves and the Rank-N plots are given in Fig. 8, for the four methods and the UBIRIS.v2 and FRGC sets. In all cases, the proposed method[1] solidly outperformed its

[1]MATLAB® source available at http://www.di.ubi.pt/~hugomcp/DeepPeriocular.zip

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2017.2771230, IEEE Transactions on Information Forensics and Security

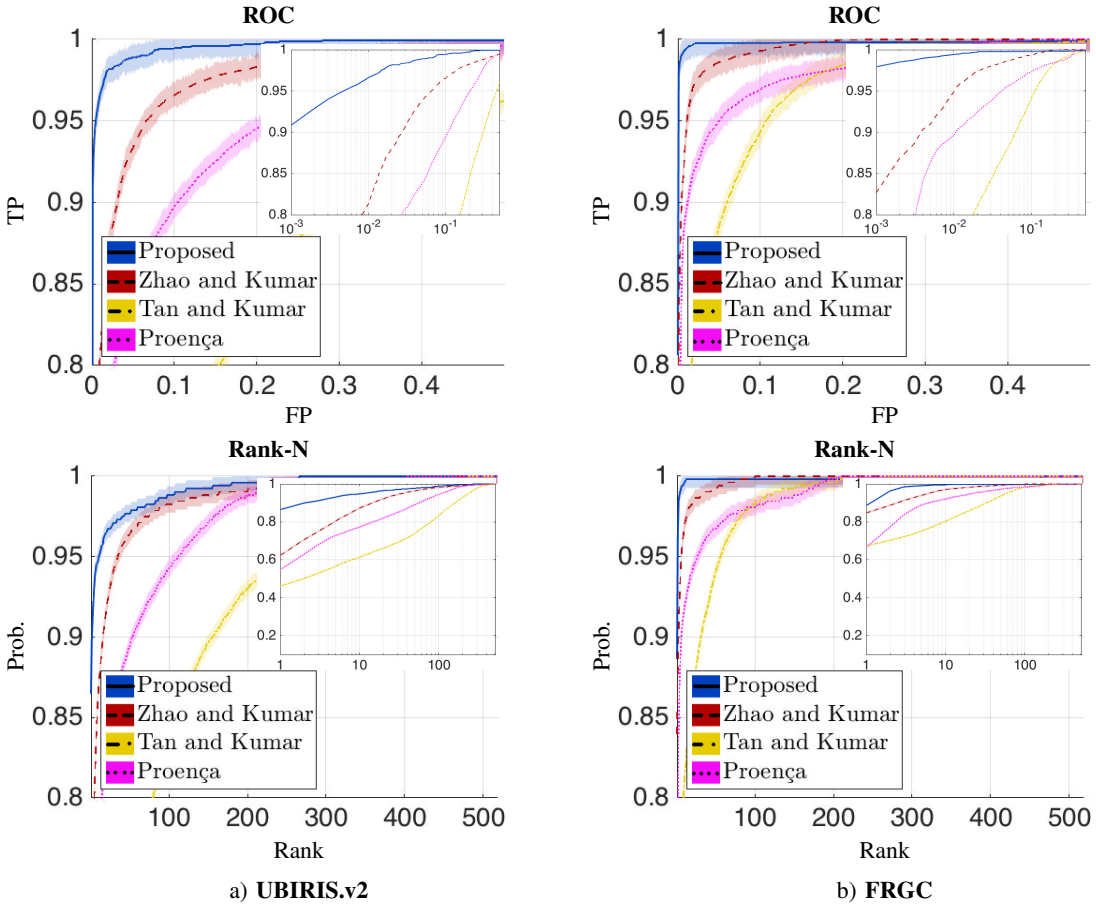

Fig. 8. Comparison between the performance attained by the method proposed in this paper and three baselines that represent the state-of-the-art. Results are given for the full UBIRIS.v2 and FRGC data sets, i.e., without disregarding any sample of these sets.

competitors, with solid differences in performance with respect to any other strategy. The differences in performance are particularly evident for small levels of false acceptances, which is exactly the most valuable operating range for security applications. Regarding the UBIRIS.v2, the proposed method attained EERs around 1.9%, decreasing the state-of-the-art rate over 80%, and over 88% in terms of Rank-1 accuracy. Results observed for the FRGC set were substantially better than those for UBIRIS.v2, which accords the previous research (e.g., [1]) and were justified by the lower number of degradation factor in this set (essentially blur and poor resolution). Again, the proposed method got the best performance among its competitors, with the true identity being reported at the first position (Rank-1) over 92% of the times. In all performance measurements, the differences with respect to the second best method (Zhao and Kumar [27]) were evident, particularly in the most important range of the performance space (FAR values less than 10^{-2}). Table I summarizes the performance indicators observed in our experiments, for the four algorithms and two data sets considered.

Method	AUC	Rank-1	EER
UBIRIS.v2			
Proposed (*Periocular* CNN)	**0.998** $\pm$ **4**e^{-4}	**0.88** $\pm$ **0.02**	**0.019** $\pm$ **6**e^{-4}
Proposed (*All* CNN)	0.994 $\pm$ 4e^{-4}	0.84 $\pm$ 0.02	0.039 $\pm$ 8e^{-4}
Zhao and Kumar [27]	0.984 $\pm$ 5e^{-4}	0.62 $\pm$ 0.02	0.109 $\pm$ 2e^{-3}
Tan and Kumar [24]	0.913 $\pm$ 3e^{-3}	0.44 $\pm$ 0.02	0.153 $\pm$ 3e^{-3}
Proença [19]	0.965 $\pm$ 1e^{-3}	0.58 $\pm$ 0.03	0.114 $\pm$ 3e^{-3}
FRGC			
Proposed (*Periocular* CNN)	**0.999** $\pm$ **4**e^{-4}	**0.92** $\pm$ **0.01**	**0.011** $\pm$ **3**e^{-4}
Proposed (*All* CNN)	0.996 $\pm$ 4e^{-4}	0.89 $\pm$ 0.02	0.028 $\pm$ 3e^{-4}
Zhao and Kumar [27]	0.995 $\pm$ 4e^{-4}	0.82 $\pm$ 0.03	0.040 $\pm$ 1e^{-3}
Tan and Kumar [24]	0.971 $\pm$ 3e^{-3}	0.69 $\pm$ 0.02	0.062 $\pm$ 2e^{-3}
Proença [19]	0.979 $\pm$ 2e^{-3}	0.70 $\pm$ 0.03	0.058 $\pm$ 3e^{-3}

TABLE I
COMPARISON BETWEEN THE PERFORMANCE OBTAINED BY THE METHOD PROPOSED IN THIS PAPER WITH RESPECT TO THREE STATE-OF-THE-ART STRATEGIES.

F. Improvements and Further Work

As insight for further improvements, Fig. 11 illustrates the samples where the proposed method obtained its worst results in terms of the Rank-n positions (UBIRIS.v2). In most cases, failures were due to: 1) large differences in phase (when the eye centre is deviated from the image centre); and 2) cropped eye regions that are too *narrow*, when the eyebrows and the skin are not available. In such cases, images contain

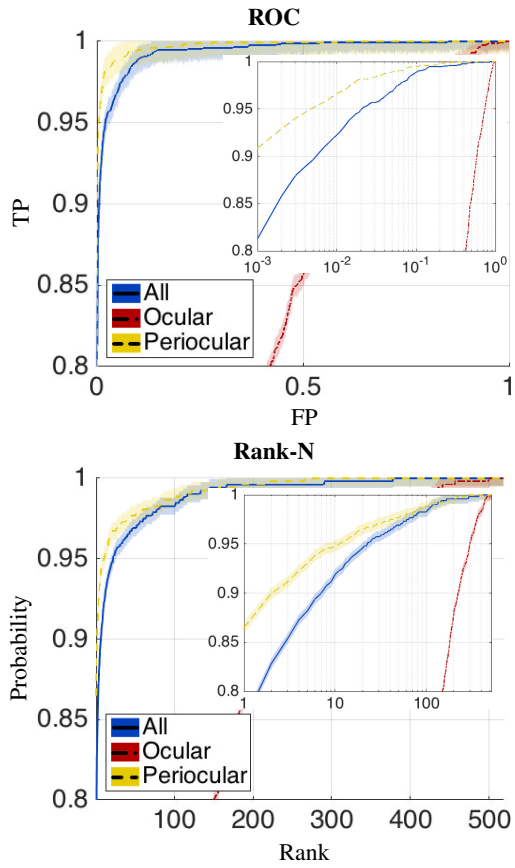

Fig. 9. Comparison between the recognition performance obtained by the CNNs when using all the information available (*All* series, represented by blue lines), when discarding the components inside the ocular globe (*Periocular* series, represented by yellow lines), and when considering exclusively the components in the ocular globe (*Ocular* series, represented by red lines).

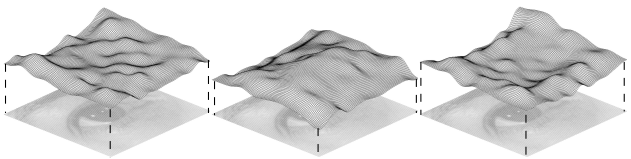

Fig. 10. Comparison between the average magnitude of the 512 (9×12) CNN filters learned immediately before the fully connected layers, i.e., the first point in the CNN where the filters coefficients have a bijective correspondence to input image regions (interpolated 45×60 grids are shown, for visualization purposes). Here, the filter magnitude corresponds directly the relevancy of the corresponding regions in the input data. Results regard the UBIRIS.v2 set and are identical to the observed for the FRGC data (not included to avoid redundancy).

almost exclusively the ocular regions, which - considering that our method disregards such information - justifies its poor performance. These problems can be attenuated if more accurate eye detection modules are used, or by considering (in a way similar to the work of Zhao and Kumar [27]) semantic information about the *narrowness* of the detected eyes, in which the narrowest samples (containing almost exclusively the ocular part) can be classified by a CNN that also considers the ocular components (corresponding to the *All* configuration results given in Sec. IV-D). Even though this network got worse performance than its *Periocular* counterpart, the performance in those narrowest samples was typically the best among all methods tested.

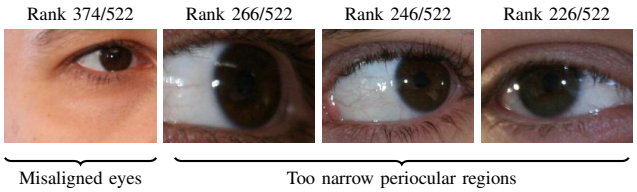

Fig. 11. Examples of the UBIRIS.v2 images where the proposed method got its worst performance. Two major error sources were detected: 1) eyes misaligned with the image centers; and 2) cases where the skin and eyebrows are badly visible.

V. Conclusions

This paper describes a periocular recognition algorithm for visible-light data that is based in convolution neural networks (CNNs). The novelty is that, by augmenting the learning data using multi-class artificial samples, it is possible to implicitly transmit prior information to the network about the regions in the input data that are not reliable for biometric recognition. Such conclusion, if left to be autonomously drew by the CNN would require additional amounts of learning data, which might not be available.

With respect to the periocular biometrics domain, there are two important conclusions: 1) for visible-light data, performance improves when the information in the ocular globe is disregarded, and the recogniser's response is solely based in the surrounding eye's components; and 2) disregarding the iris/sclera regions can be done without explicitly segmenting these regions during the recognition step. As main result, the proposed method advances the state-of-the-art performance in the *closed-world* scenario for two of the most used data sets in this field (UBIRIS.v2 and FRGC). It should be noted that these results were observed when considering even the poorest quality samples in both data sets, i.e., without disregarding any image or using any *friendly* versions of the datasets.

Acknowledgements

We acknowledge the support of *NVIDIA Corporation*®, with the donation of one *Titan X GPU*.

This work was supported by PEst-OE/EEI/LA0008/2013 research program.

References

[1] F. Alonso-Fernandez and J. Bigun. A survey on periocular biometrics research. *Pattern Recognition Letters*, vol. 82, pag. 92–105, 2016.

[2] S. Bharadwaj, H. Bhatt, M Vatsa and R. Singh. Periocular biometrics: When iris recognition fails. In Proceedings of the IEEE International conference on Biometrics: Theory, Applications and Systems, doi: 10.1109/BTAS.2010.5634498, 2010.

[3] R. Byrd, M. Hribar and J. Nocedal. An Interior Point Algorithm for Large-Scale Nonlinear Programming. *SIAM Journal on Optimization*, vol. 9, no. 4, pag. 877–900, 1999.

[4] Z. Cao and N. Schmid. Fusion of operators for heterogeneous periocular recognition at varying ranges. *Pattern Recognition Letters*, vol. 82, pag. 170–180, 2016.

[5] L. Chen and J. Ferryman. A Comparative Analysis of Two Approaches to Periocular Recognition in Mobile Scenarios. In Proceedings of the IEEE International conference on Biometrics: Theory, Applications and Systems, pag. 1–6, doi: 10.1109/BTAS.2015.7358753, 2015.

[6] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feed forward neural networks. In Proceedings of the International Conference on Artificial Intelligence and Statistics, pag. 249–256, 2010.

[7] K. Hollingsworth, K. W. Bowyer and P. Flynn. Improved Iris Recognition Through Fusion of Hamming Distance and Fragile Bit Distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pag. 2465–2476, 2011.

[8] A. Joshi, A. Gangwar, R. Sharma, A. Singh and Z. Saquib. Periocular recognition based on Gabor and Parzen PNN. In Proceedings of the IEEE International Conference on Image Processing, doi: 10.1109/ICIP. 2014.7026008, 2014.

[9] A. Krizhevsky, I. Sutskever and G. Hinton. *Imagenet* classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems Conference, pag. 1097–1105, 2012.

[10] J. Long, E. Schelhamar and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pag. 640–651, 2015.

[11] G. Mahalingam and K. Ricanek Jr. LBP-based periocular recognition on challenging face datasets. *EURASIP Journal on Image and Video Processing*, vol. 36, pag. 1–13, 2013.

[12] L. Nie, A. Kumar and S. Zhan. Periocular Recognition using Unsupervised Convolutional RBM Feature Learning. In Proceedings of the 22^{nd} International Conference on Pattern Recognition, pag. 399–404, 2014.

[13] I. Nigam, M. Vatsa and R. Singh. Ocular biometrics: A survey of modalities and fusion approaches. *Information Fusion*, vol. 26, pag. 1–35, 2015.

[14] U. Park, A. Ross and A. Jain. Periocular Biometrics in the Visible Spectrum: A Feasibility Study. In Proceedings of the 3^{rd} IEEE International conference on Biometrics: Theory, Applications and Systems, pag. 153–158, 2009.

[15] U. Park, R. Jillela, A. Ross and A. Jain. Periocular Biometrics in the Visible Spectrum. *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 1, pag. 96–106, 2011.

[16] P.J. Phillips, P. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the Face Recognition Grand Challenge. In Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pag. 947–954, 2005.

[17] H. Proença. Iris recognition: On the segmentation of degraded images acquired in the visible wavelength. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pag. 1502–1516, 2010.

[18] 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, pag. 1529–1535, 2010.

[19] H. Proença. Ocular Biometrics by Score-Level Fusion of Disparate Experts. *IEEE Transactions on Image Processing*, vol. 23, no. 12, pag. 5081–5093, 2014.

[20] R. Raghavendra, K. Raja, B. Yang and C. Busch. Combining Iris and Periocular Recognition Using Light Field Camera. In Proceedings of the IAPR Asian Conference on Pattern Recognition, doi: 10.1109/ACPR. 2013.22, 2013.

[21] A. Ross, R. Jilela, J. Smereka, V. Boddeti, B. Kumar, R. Barnard, X. Hu, P. Pauca and R. Plemmons. Matching Highly Non-ideal Ocular Images: An Information Fusion Approach. In Proceedings of the 5^{th} IAPR International Conference on Biometrics, doi: 10.1109/ICB.2012.6199791, 2012.

[22] A. Sequeira, L. Chen, J. Ferryman, F. Alonso-Fernandez, J. Bigun, K. Raja, R. Raghavendra, C. Busch and P. Wild. Cross-Eyed - Cross-Spectral Iris/Periocular Recognition Database and Competition. In Proceedings of the 2016 International Conference of the Biometrics Special Interest Group (BIOSIG), doi: 10.1109/BIOSIG.2016.7736915, 2016.

[23] C. Szegedy, A. Toshev and D. Erhan. Deep Neural Networks for Object Detection. In Proceedings of the Advances in Neural Information Processing Systems Conference, pag. 2553–2561, 2013.

[24] C. Tan and A. Kumar. Towards Online Iris and Periocular Recognition Under Relaxed Imaging Constraints *IEEE Transactions on Image Processing*, vol. 22, no. 10, pag. 3751–3765, 2013.

[25] A. Vedaldi and K. Lenc. MatConvNet – Convolutional Neural Networks for MATLAB. In Proceedings of the 23^{rd} ACM International Conference on Multimedia, pag. 689–692, 2015.

[26] D. Woodard, S. Pundlik, J. Lyle and P. Miller. Periocular Region Appearance Cues for Biometric Identification. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition Workshops , doi: 10.1109/CVPRW.2010.5544621, 2010.

[27] Z. Zhao and A. Kumar. Accurate Periocular Recognition under Less Constrained Environment Using Semantics-Assisted Convolutional Neural Network. *IEEE Transactions on Information Forensics and Security*, doi: 10.1109/TIFS.2016.2636093, 2016.