A Deep Adversarial Framework for Visually Explainable Periocular Recognition

Anonymous CVPRW 2021 submission

Paper ID ****

Abstract

In the biometrics context, the ability to provide the reasoning behind a decision has been at the core of major research efforts. Explanations serve not only to increase the trust amongst the users of a system, but also to augment the system's overall accountability and transparency. In this work, we describe a periocular recognition framework that not only performs biometric recognition, but also provides visual representations of the features/regions that supported a decision. Being particularly designed to explain non-match ("impostors*") decisions, our solution uses adversarial generative techniques to synthesise a large set of "*genuine*" image pairs, from where the most similar elements with respect to a query are retrieved. Then, assuming the alignment between the query/retrieved pairs, the element-wise differences between the query and a weighted average of the retrieved elements yields a visual explanation of the regions in the query pair that* would have to be different *to transform it into a "*genuine*" pair. Our quantitative and qualitative experiments validate the proposed solution, yielding recognition rates that are similar to the state-of-the-art, but - most importantly - also providing the visual explanations for every decision.*

1. Introduction

This work describes an integrated framework for periocular biometric recognition which - apart performing the recognition task - also provides a visual explanation that sustains every decision. Considering the biometric recognition ubiquity and dependability [2], our main goal in this paper is not to propose a *better* recognition framework in terms of the error rates, but to particularly diverge of the black-box paradigm and follow a *visually explainable* paradigm, as illustrated in Fig. 1.

[0]The code is publicly available at http://github.com/anonymized

Figure 1. Key insight of the proposed visual explainable framework: given a pair of images, the system **not only reports a binary decision** ("*genuine*"/"*impostor*" classes), but also **highlights the regions in each sample that contributed the most in case of a *non-match* decision**. In this example, yet the iris and skin colour are similar between samples, the eyebrows and eyelashes shapes are evidently different, along with a skin spot in the sample illustrated at the left side. These are exactly the regions highlighted in the visual explanations.

Typically, a recognition problem involves a set of unique and non-transferable features that can unmistakably identify a subject. Biometric traits, as they are designated in the field, serve such purposes, as long as they are universal, distinguishable, resilient to changes and easy to collect [1]. Upon proving their compliance with these requirements, biometric traits can be divided into two major categories:

1. *Physiological* features (e.g., the iris, fingerprint and retina) that are naturally possessed by a given subject;

2. *Behavioural* biometrics, that yield from the interaction between a subject and the surrounding environment (e.g., the gait and handwritten signature) [18].

Concentrating growing interests in the biometrics domain, periocular recognition uses the information in the

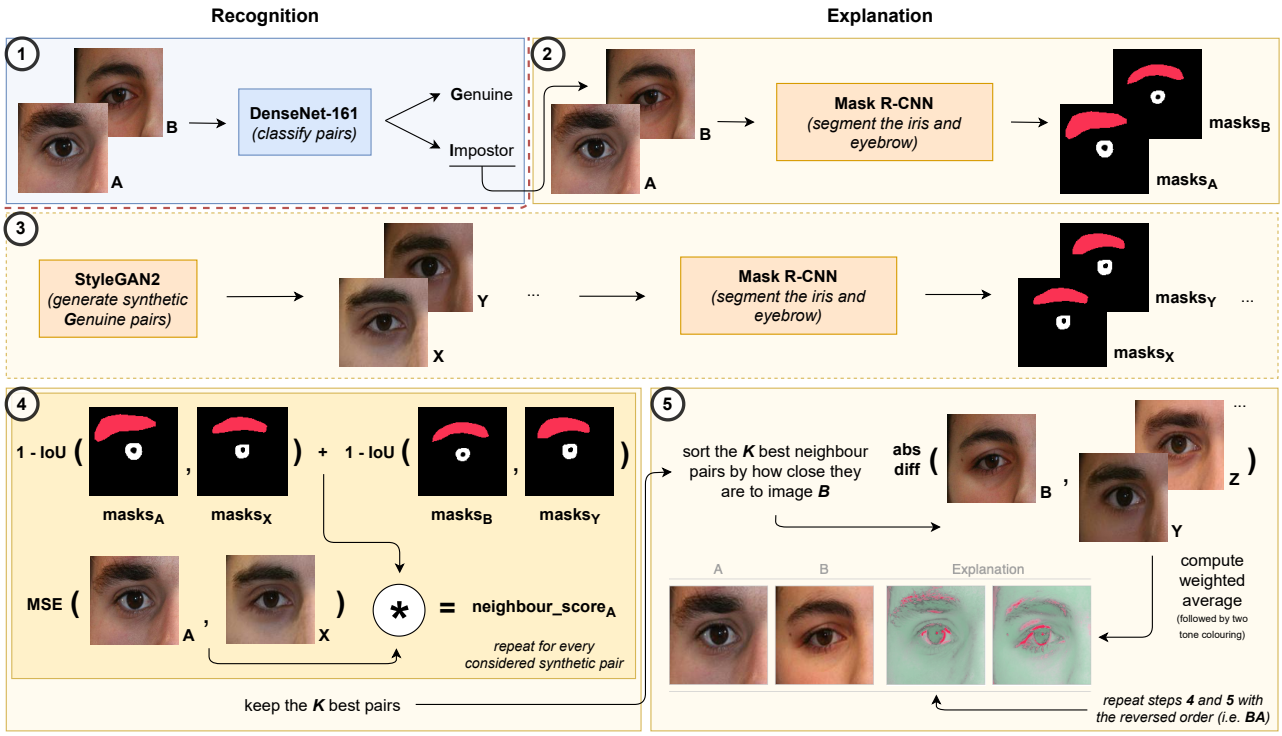

Figure 2. Cohesive perspective of the main pipeline of the proposed solution. The first step (recognition) encompasses a CNN that distinguishes between "*genuine*" and "*impostor*" pairs. Then, upon an "*impostor*" decision, steps two to five (explanation) find the K "*genuine*" synthetic pairs among a large set that most closely resemble the query pair. Assuming the alignment between the query and the retrieved pairs, the element-wise differences between the query and a weighted average of the retrieved elements provides a visual explanation of the regions/features in the query that would have to be different to turn the query into a "*genuine*" pair.

vicinity of the eye to perform recognition, in which the iris, sclera, eyebrow, eyelid and skin stand out.

Regarding the concept of *explainability* and its application to recognition problems, it should be noted that Deep Learning solutions rely on model complexity and abstraction prowess to become truly accurate. Although seemingly innocuous, there could be seriously negative outcomes if such *black-boxes* gamble on the clearance of unauthorised people into sensible areas. Hence, it is particularly important to provide human understandable explanations of the decisions, which will augment the overall system accountability and transparency, enabling a broader range of applications (i.e., forensics). Recently, the EU, through the GDPR [5], introduced the notion of "*right to an explanation*". Even though the definition and scope of such explanations are still subject to debate [20], these are definite strides towards a formal regulation that will surely augment the importance given to the concept of explainability.

According to the above points, this paper describes a framework that receives a pair of images and returns a two-fold output: 1) a binary match/non-match decision, that discriminates between the "*genuine*"/"*impostor*" pairs; and 2) a *visual explanation* that highlights the features/regions of the input data that sustained a particular decision. This is considered the main contribution of our work, in the sense that - to the best of our knowledge - it is the first that creates an accurate and explainable representation of the reasons behind certain decisions of the recognition system. Other contributions include the use of Generative Adversarial Networks (GANs), to synthesise visually pleasant images pair that faithfully resemble the distribution of the "*genuine*" pairs, which augments the variety and flexibility of the learning set and can be seen as an alternate form of data augmentation.

Fig. 2 provides a cohesive overview of the framework that performs the recognition task and provides the corresponding explanations: 1) at first, a CNN (of a well known architecture) is trained to discriminate between match/non-match decisions. If the pair is deemed to belong to the "*impostors*" distribution, we find its most similar "*genuine*" pairs in a large set of synthetic data. The insight here is that, even if the query pair has significant differences between its elements that led to an "*impostor*" decision, the closest synthetic pairs most likely do not (as they were drawn from

the "*genuine*" distribution). Then, assuming that the most likely synthetic pairs are sufficiently aligned to the query, by obtaining the pixel-wise weighted differences between the query and its K closest neighbours, the data disparities become evident.

The remainder of this paper is organised as follows: Section 2 summarises the most relevant research in the fields of periocular recognition and Machine Learning Explainability. Section 3 describes our method and Section 4 analyses the results obtained. Section 5 concludes this paper, while also providing some final remarks.

2. Related Work

2.1. Periocular Recognition

The seminal breakthroughs in the periocular recognition problem can be traced to a set of methods termed *feature descriptors*. Methods such as HoG, LBP and SIFT were able to produce simplified data representations by relying on edges, textures and keypoints, respectively. In [22], the results from each feature descriptor were fused to faithfully discriminate between the "*genuine*"/"*impostor*" pairs. This work served as basis for subsequent fusion-based approaches, as in [6]. In [11] a Restricted Boltzmann Machine was used to learn a probabilistic distribution over the input data, further discriminated with metric learning and SVMs.

With the effective application of Deep Learning solutions, researchers turned to popular architectures (in particular Convolutional Neural Networks), to pursuit ever increasing recognition accuracy. Accordingly, in [23] the main concept involves the use of multiple CNNs that are specialised in classifying a particular kind of semantic information (e.g. gender or age). Then, a score fusion process yields the final response. In [17], authors enforce a CNN to ignore the ocular region (due to its likelihood to contain specular reflections) and rely in the eye's surrounding area (eyebrow, eyelid and skin). [19] created independent representations of the iris and periocular regions, that feed classification modules, whose scores are finally fused to reach the decision. Using a multi-glance mechanism, where part of the intermediate components are configured to incorporate emphasis on the most important semantical regions (i.e., eyebrow and eye), Zhao and Kumar [24] developed a recognition model that particularly focus these regions, enabling the deep Convolutional Neural Network (CNN) to learn additional discriminative features that improve the recognition capability of the whole model. Recently, [4] attempted to bridge the gap between biometric recognition and interpretability, by learning feature specific filters that respond to a range of preferred spatial locations. [9] propose an integrated solution that leverages the discovery of parts as a form of attention.

2.2. Machine Learning Explainability

In the literature, the existing explainable techniques are commonly divided in terms of their depth, scope and model applicability [12], [15]. Depth is related to the length to which we explain a given model, i.e. whether the technique limits the model's complexity to make it more transparent (*intrinsic* explainability) or allows complexity and focuses on explaining exclusively the system outputs (*post hoc* explainability). Scope indicates the range that a technique possesses, i.e., if it explains individual predictions (*local*) or the model's entire behaviour (*global*). Finally, applicability divides the techniques based on their model affinity, i.e. whether they are only compatible with a specific family of models (*model-specific*) or any kind of model (*model-agnostic*). The most commonly cited techniques include LIME [14] and Shapley codes (SHAP) [13]. The former uses a surrogate linear model, trained on perturbed data (e.g. disabled clusters of adjacent pixels), to locally approximate the behaviour of a complex black-box model. The latter uses game theory and Shapley values, which are assigned to the features based on how important they are to a given prediction. Additionally, Saliency Maps [10] use the derivative of a highly complex function (essentially, a CNN) with respect to a given input image, to determine which pixels need to be changed the least, while also changing the output class the most. Finally, for visualisation purposes and, therefore, outside the scope of this work, PDP [7] and ALE [3] techniques are able to produce plots that correlate the independent variables to a target variable, exploiting the notions of marginal and conditional distributions, respectively.

3. Proposed Method

3.1. Learning Phase

The main components of the proposed method comprise three well known models: the DenseNet-161, Mask R-CNN and StyleGAN2. The first one (DenseNet-161) is trained to solve an identity verification problem, while the segmentation model (Mask R-CNN) is fine-tuned to produce high-quality masks for the iris and eyebrow. Finally, the GAN model (StyleGAN2) learns how to create synthetic data that, while closely resembling the distributions in the training set, is diverse enough to approximate unseen subjects. Additionally, a fourth, auxiliary model (ResNet-18) is fitted to discriminate between images from the left and right sides of the face. Although trained separately, all the models learn from the same training split, which excludes a set of disjoint IDs that are reserved for performance evaluation purposes.

Regarding the model used in the verification task (DenseNet-161), it should be stated that it has mach more parameters than the network used by Zhao and Kumar [23] in their solution. This might be the fact that sustained slightly better recognition performance of our model with

respect to the baseline (Sec. 4.3), but also at the expense of a substantial higher computational cost of classification than the baseline, which might be impracticable in some cases.

3.2. Inference Phase

Once trained, our method is conceptually divided into five major steps, as depicted in Fig. 2. Firstly, the DenseNet-161 model is used to verify the claimed identity: upon receiving a pair of images, the model discriminates between "*genuine*"/"*impostor*" pairs. If the pair is deemed to be "*impostor*", the remaining steps create a visually interpretable explanation of that decision.

The second step takes the query pair and, using Mask R-CNN, segments the irises and eyebrows regions. Next, step three uses the StyleGAN2 generator to create a large, synthetic set of exclusively "*genuine*" pairs (i.e. where both images belong to the same person). For each of these synthetic pairs, the ResNet-18 model determines its side configuration (i.e. whether images regard the left or right side of the face) and, as before, masks are obtained by the segmentation model.

After obtaining the synthetic data and their corresponding masks, the synthetic dataset is indexed based on the coordinates of the center of the iris, which will enable faster search in the retrieval step. To that end, the clustering algorithm K-Means is trained on a subset of the iris segmentation masks to obtain three centroids, one for each major iris gaze family (i.e. left, centre and right). This way, we index the available pairs based on their combination of iris positions (e.g. left-left, right-centre . . .). By doing so, when searching, we can just rely on the synthetic pairs that share the same combination as the test pair, saving time and useless calculations.

Upon settling for a portion of the synthetic dataset that closely meets the iris position constraint, the segmentation masks are further used to determine which synthetic pairs have the iris and eyebrow approximately overlapped to the query. This is an important requirement to obtain visually pleasant explanations, given that pixel-wise differences are extremely sensitive to differences in phase (i.e., component misalignment). Accordingly, we obtain a similarity score s_X between each synthetic neighbour and the query, given by:

$$s_X = \omega_{\text{masks}} * ||\text{query\_pair}_\text{A} - \text{neighbour}_\text{X}||_2, \quad (1)$$

being $||.||_2$ the $\ell - 2$ norm and $\omega_{.}$ a weight that considers component misalignment. This way, we obtain a weighted distance between each synthetic neighbour and the first image of the query pair. ω_{masks} values serve to favour pairs that have good alignment, considering $1 - \text{IoU}(.,.)$, i.e., the complement of the intersection-over-union of the synthetic/query segmentation masks. In practice, we search amongst the (large) thousands of synthetic pairs, the closest to the query pair in terms of the first image. Therefore, given that the second image of the query pair is from a different subject, it will most likely have features that are different to the synthetic neighbours, which are exactly the kind of dissimilarities that make up the final explanations.

This way, the K closest neighbours are sorted according to their element-wise distance to image B, using (2). Finally, to produce the final explanation, the K best neighbours are used to obtain the pixel-wise differences against the query pair image B. In practice, a neighbour distance is subtracted from the total sum of distances, creating an inverted distance. This assures that the contribution of the closest synthetic neighbours to the final result is more important than of those with bigger distances.

3.3. Implementation Details

The DenseNet-161 model was trained for 15 epochs with a learning rate of 0.0002 and a batch size of 64 image pairs. The Adam algorithm was used for the weight optimisation process (with default β_1 and β_2 values). A similar training setup was used to train the ResNet-18 model, albeit for a smaller number of epochs (i.e. 5). For the Mask R-CNN's training process, we kept its default values, using a learning rate of 0.001, a batch size of 1 and 30 epochs worth of training (in this case, fine-tuning from the COCO pre-trained weights). Regarding the StyleGAN2 architecture, the used training step comprised a total of 80.000 iterations and a batch size of 8. After converging, the generator is capable of synthesising realistic looking images, such as the roughly 400.000 pairs that make up the artificial dataset. Finally, for the number K, that determines how many synthetic neighbours should be kept, we used a default value of 15.

4. Experiments and Discussion

4.1. Datasets and Working Scenario

As mentioned above, the proposed framework is composed of two modules: 1) one for recognition; and 2) the other for explanation purposes. Regarding the former, the chosen CNN is solely trained on the UBIPr dataset [16], which provides the ID annotations used in the identity verification problem. Regarding the explanation step, it mainly relies on a combination of UBIPr and FFHQ [21]. Despite not being directly applicable to the context of this work (i.e. it contains full face images, thus requiring extra steps to extract the periocular region), the FFHQ dataset contains a large variety in terms of periocular attributes, some of which are scarcer in the UBIPr dataset. In practice, a small, but curated, portion of the FFHQ samples was used to cre-

ate a data super set (Fig. 3). Regardless of their source, all images were resized to a common shape, depending on the task (i.e. 512x512x3 for Mask R-CNN, 256x256x3 for StyleGAN2 and 128x128x3 for the CNNs).

Figure 3. Samples from the two datasets used. The top row represents images of the UBIPr dataset, whereas the bottom row illustrates cropped samples of the FFHQ dataset.

As it is usual in the biometric recognition context, it is important to define proper working modes and world settings, for which the system is built. With respect to the working mode, our model runs in verification mode (also referred to as *one-to-one*), where the system validates a claimed identity [1]. As for the world setting, we assume an open-world setting, meaning that unseen subjects can be faithfully handled in the inference step.

4.2. Explainability Evaluation

Our explainability chain starts by the train of a DenseNet-121 model to perform the verification task. This model can be further paired to either LIME, SHAP or Saliency Maps to create comprehensive comparison schemes, to which we add the method described in [9]. Fig. 5 provides several examples of the synthetic "*genuine*" images pairs generated from the GAN model. Apart their obvious visual realism, it is important that this set contains samples with the most likely known data covariates for the periocular region: varying gazes, wide-opened/closed eyes, varying poses, partial occlusions, and even varying facial expressions. Failing in incorporating such diversity will determine that the closets synthetic pairs of a query will still be notoriously different from it, and that the visual representations obtained will have poor realism.

Fig. 4 displays the expected results from a visually explainable system. In practice, LIME tries to keep the most important super-pixels, SHAP highlights those it deems important in red tones and Saliency Maps produce greyscale explanations. As for the method by Huang and Li, it generates a heat-map in which red tones elevate important areas. Focusing on the common pairs between all methods, the left sample is essentially different with regards to eyebrow thickness and presence/absence of a noticeable skin spot. As for the right one, the most obvious disparities have to do with the eyebrow areas. Overall, our results are the most informative, when compared with the remaining four solutions. While LIME and SHAP do a decent job, Saliency Maps provide a faint explanation. It is Huang and Li's method that comes closer to our level of visual appeal, by clearly highlighting portions of the eyebrow and a portion of subject A's skin spot, in the left pair. Moreover, when given the right sample, it generates a solid red area comprising subject B's eyebrow. However, upon closer inspection, our results show more appealing visual cues: in the left sample, distinct red tones on top of A's skin spot and eyelashes, as well as, reiterated eyebrow differences in the right sample with highlights in both eyebrows, rather than just one. As for the remaining samples, the third (just below the first) is clearly explained by highlighting the entirety of both skin areas, which are obviously different between images A and B. Finally, in the fourth pair it is also shown how the eyelids differ, by colouring that periocular component on subject B's image, and, in the fifth sample, subjects B's eyebrow and iris are accurately shown in red.

When objectively measuring the differences between the explanations provided by the proposed method and the baselines (LIME, SHAP, Huang and Li (HL) and Saliency Maps (SM)), we used a set of 10 heterogeneous test queries and measured the pixel-wise explanation coefficients returned by each technique, which correspond to the importance (weight) given by each method to a particular image position for a decision. Next, considering that any meaningful correlations between the responses of two methods would have to be linear, we measured the Pearson's linear correlation between pairs of techniques:

$$r_{xy} = \frac{\sum_i (x_i - \hat{x})(y_i - \hat{y})}{\sqrt{\sum_i (x_i - \hat{x})^2 \sum_i (y_i - \hat{y})^2}}, \qquad (2)$$

where x_i/y_i denote the ith scores provided by each technique and the $\hat{}$ symbol denotes the mean value. This way r_{xy} measures how similar are the explanations provided by the x and y techniques: values close to 0 will correspond to more independent explanations, while values towards one will point for semantic similarities between the explanations provided by both techniques.

The results are provided in the confusion matrices shown at Fig. 6, where the main diagonal provides the distributions of the scores generated by each technique and the remaining cells provide the scatter plots between pairs of techniques with the Pearson's correlation value r_{xy} given at the top left corner of each cell ('SM' stand for Saliency Maps and 'HL' denotes the Huang and Li solution)). All these techhniques report a local numeric value that corresponds to the role/importance of each region in the final decision. The exception is LIME, where the pixels are binary discriminated into "visible"/"occludded'. In this case, we considered that "visible" will be equal to 1, while 'occludded" will be equal to 0. Overall, we observed that the techniques provide relatively independent responses for the

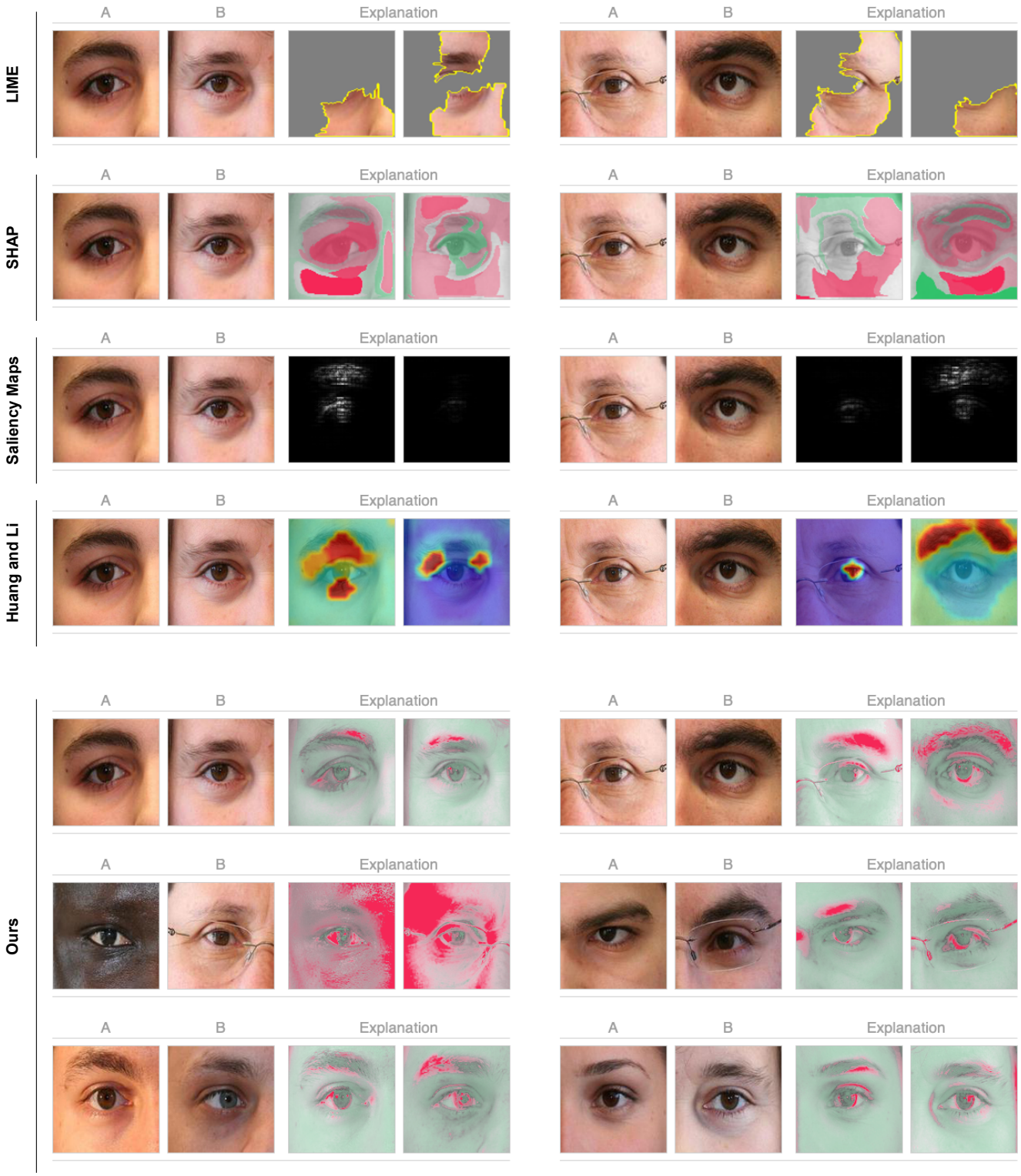

Figure 4. Examples of the results attained by three standard interpretability techniques (LIME, SHAP and Saliency Maps), a state-of-the-art interpretable deep model for fine-grained visual recognition (i.e. [9]) and our method. Notice how our results are clearer in highlighting the components that justify every non-match decision (e.g., skin texture and color eyebrows/eyelashes size and distribution, irises color and even skin spots).

Figure 5. Examples of the synthetic image pairs in our data set, generated according to a GAN model.These elements are drawn exclusively from the "genuine" distribution. Upon a query, the most similar synthetic pairs with respect to the query are found, which will provide the features/regions that would transform the query into a "*genuine*" comparison.

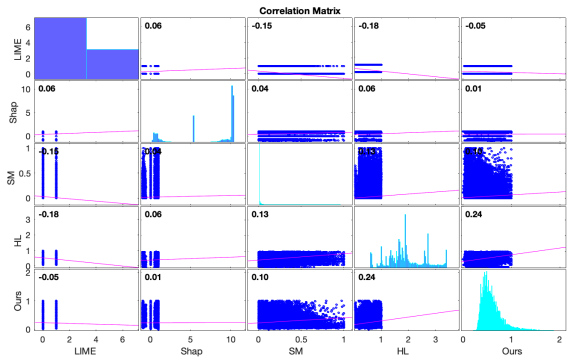

Figure 6. Pearson correlation values between the pixel-wise responses provided by the method proposed in this paper (Ours) and four baselines techniques (LIME, SHAP, Huang and Li (HL) and Saliency Maps (SM)).

importance given to each pixel in the final decision. Interestingly, in some cases, there are even negative correlation values between two methods (e.g., HL and LIME or SM and LIME). There are other pairs of solutions that achieved almost full independence between their responses (the Shapley/Ours methods), which points for completely different strategies being used to define the explaining regions/features. Still considering our method, its levels of correlation were kept relatively low with respect to the remaining methodologies, achieving values of 0.24 with respect to the method of Huang and Li (the most correlated), and 0.1 for Saliency maps. Still, we concluded that the proposed solution is extracting semantic information (e.g., features and regions) of the vicinity of the human eye that is evidently different of the kind of information emphasised by any of the remaining methods, which supports the usefulness of the solution described in this paper.

4.3. Recognition Accuracy Evaluation

At first, note that we do not aim at providing a *better* recognition framework than the state-of-the-art, in terms of the recognition rates. Even though, our main purpose in this section was to perceive if the proposed recognition/explanation network is able to achieve competitive recognition performance with respect to the state-of-the-art.

We compare the recognition effectiveness of the proposed method with respect to a well known periocular recognition model (due to Zhao and Kumar [23], considered to represent the state-of-the-art). Using the UBIRIs.v2 set [8] and the learning/evaluation protocols described in [23], we obtained the results summarised in Table 1. Also, we provide ROC values of the proposed strategy, that can be fairly combined with the similar ROC plot provided by the original authors of the baseline in [24].

A bootstrapping-like strategy was used, by sampling 90% of the available data in UBIRIS.v2 and dividing the resulting samples between two disjoint sets: 80% for training and the remaining 20% for test. The models were trained separately in each sample and the performance evaluated in the corresponding test set, from where the EER and AUC scores were obtained. This process was repeated 10 times, to perceive the mean $\pm$ standard deviations values for both metrics. Overall, results were satisfactory, particularly considering that - due to our modular design - the recognition module of the proposed framework can be easily replaced by any other, while keeping its explainability abilities.

Method	EER	AUC
Ours (open-world)	$0.108 \pm 3e{-2}$	$0.813 \pm 5e{-2}$
Ours (closed-world)	$\mathbf{0.087 \pm 2e{-2}}$	$\mathbf{0.910 \pm 2e{-2}}$
Zhao and Kumar [23]	$0.109 \pm 2e{-3}$	–

Table 1. Comparison between the recognition rates attained by the proposed method (in both world settings) and a state-of-the-art method (strictly operating in a closed-world setting). Results are given for the same learning/test sets of the UBIRIS.v2 dataset.

For reference purposes, Fig 7 provides the Receiver Operating Characteristic curves of our solution. When comparing to the corresponding results reported by authors in [23] in the same set, it can be seen a close recognition summary performance between both methods (summarised in Table 1). Overall, we observed a similar performance between both techniques in this data set, which supports the idea that the proposed solution is able to approach state-of-the-art recognition rates.

4.4. Ablation Studies

For our ablation experiments, we identified two hyperparameters of our method that might play the most signif-

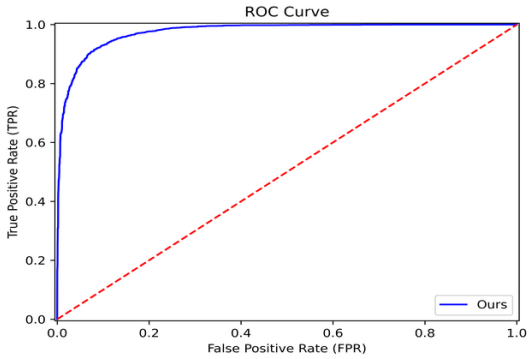

Figure 7. Receiver Operating Characteristic (ROC) curve obtained for the proposed method in data of the UBIRIS.v2 set, according to the empirical protocol designed by Zhao and Kumar [23]. The ROC curve corresponds to an EER value of about 0.087 and a AUC value of 0.910.

icant roles in the final effectiveness of the whole solution: 1) the number of neighbours retrieved (K) from the synthetic set for every query; and 2) the length of the synthetic set itself. This section discusses how changes in these values affect the quality of the generated explanations in a less than optimal way (as seen in Fig. 8).

4.4.1 Number of Neighbours

The value K determines how many synthetic pairs are considered with respect to a query. Overall, we observed that smaller values lead to more sensitive and jagged results. Up to a certain point (e.g. 15), increasing K typically enables to obtain *smoother* explanations, due to the larger number of samples taken into account when averaging the closest neighbours. This trend, however, starts returning incremental improvements (notice in Fig. 8, where $K >= 50$ progressively stops presenting a prominent tone on the eyelid).

4.4.2 Length of the Synthetic Dataset

This is the most sensitive parameter of our solution. Considering that it is important to find "*genuine*" pairs that closely resemble a query, it is particularly sensitive to assure that all typical periocular data variations are faithfully represented in the synthetic set, assuring that the retrieved elements (i.e., the most similar) will have its major components (iris, eyebrows and eyelids) aligned to the query itself. If this condition is not satisfied, the explanations loose their biological plausibility and effectiveness. Fig. 8 illustrates how smaller synthetic sets lead to less evident explanations, especially around the eyelid and the eyebrow.

5. Conclusions and Further Work

This paper described an integrated framework, based in well known deep-learning architectures, to simultaneously perform periocular recognition and - most importantly - to provide visual explanations of the regions/features that sustained every *non-match* decision, which we consider to be the cases where explanations are the most required. According to the powerful generative ability of GANs, we create a very large set of synthetic pairs that follow the "*genuine* distribution". At inference time, for every "*impostor*" comparison we are able to perceive the regions and features that *failed the most* (i.e., those that most evidently were different from a subset of the "*genuine*" (synthetic) pairs). This enables to generate pleasant visual explanations, where each component of the periocular region appears with a different colour depending on how it influenced the final decision. Importantly, the modular nature of the proposed method ensures that the periocular region can be replaced by other biometric traits (e.g., the face) without compromising the explanations.

As future work, we are developing a strategy for also providing intuitive explanations of the "*genuine*" observations, where the strategy has to be very different from the idea behind the "*impostors*" insight used in this paper.

Acknowledgements

This work is funded by FCT/MEC through national funds and co-funded by FEDER - PT2020 partnership agreement under the project UIDB//50008/2020. Also, it was supported by operation Centro-01-0145-FEDER-000019 - C4 - Centro de Competências em Cloud Computing, co-funded by the European Regional Development Fund (ERDF) through the Programa Operacional Regional do Centro (Centro 2020), in the scope of the Sistema de Apoio a Investigacão Cientifica e Tecnologica - Programas Integrados de IC&DT.

References

[1] A. Ross A. K. Jain and S. Prabhakar. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):4–20, 2004. 1, 5

[2] D. Braines R. Tomsett A. Preece, D. Harborne and S. Chakraborty. Stakeholders in explainable ai. arXiv:1810.00184 [cs], Set. 2018. Accessed on: Feb. 27, 2021. [Online]. Available: http://arxiv.org/abs/1810.00184. 1

[3] D. W. Apley and J. Zhu. Visualizing the effects of predictor variables in black box supervised learning models. arXiv:1612.08468 [stat], Ago. 2019. Accessed on: Feb. 27, 2021. [Online]. Available: http://arxiv.org/abs/1612.08468. 3

[4] H. Li X. Shen B. Yin, L. Tran and X. Liu. Towards interpretable face recognition. *2019 IEEE/CVF International*

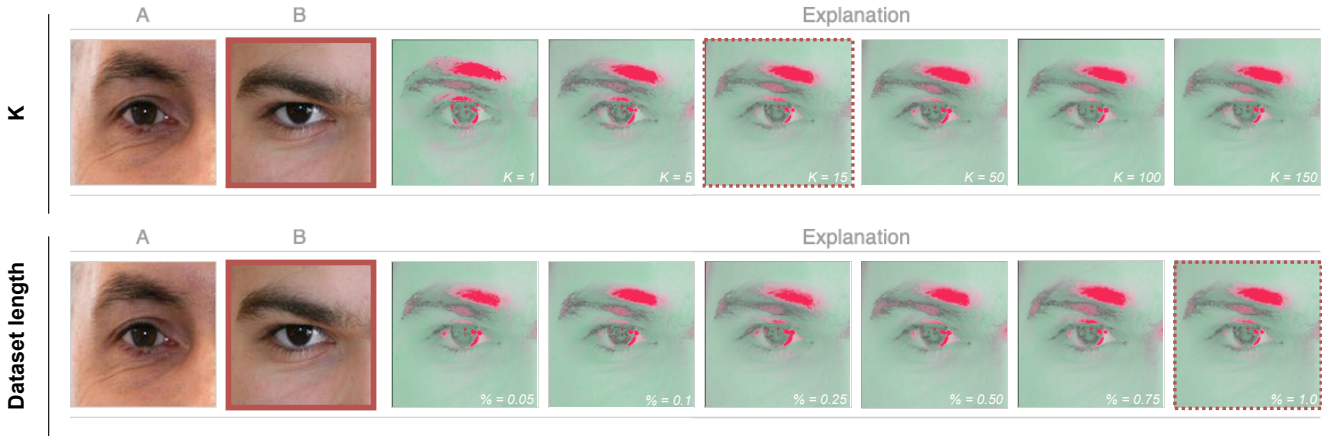

Figure 8. Typical changes in the results when the two most important parameters of the proposed method are varied. The red square indicates which image is being *explained* (i.e. B), while the red dashed squares provide the default values used in our experiments. In general, increasing K up to 15 allows for smoother explanations, as does keeping a large dataset. Reducing the latter tends to produce less sensitive results, substantially decreasing the plausibility of the visual explanations generated.

Conference on Computer Vision (ICCV), pages 9347–9356, 2019. 3

[5] European Commission. General data protection regulation. 2018. Accessed on: Feb. 27, 2021. [Online]. Available: https://gdpr-info.eu. 2

[6] A. Ross et al. Matching highly non-ideal ocular images: An information fusion approach. *2012 5th IAPR International Conference on Biometrics (ICB)*, pages 446–453, 2012. 3

[7] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. 3

[8] R. Santos J. Oliveira H. Proença, S. Filipe 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*, 32(8):1529–1535, 2010. 7

[9] Z. Huang and Y. Li. Interpretable and accurate fine-grained recognition via region grouping. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8659–8669, 2020. 3, 5, 6

[10] A. Vedaldi K. Simonyan and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv:1312.6034 [cs], Abr. 2014. Accessed on: Feb. 27, 2021. [Online]. Available: http://arxiv.org/abs/1312.6034. 3

[11] A. Kumar L. Nie and S. Zhan. Periocular recognition using unsupervised convolutional rbm feature learning. *2014 22nd International Conference on Pattern Recognition*, pages 399–404, 2014. 3

[12] Z. C. Lipton. The mythos of model interpretability. arXiv:1606.03490 [cs, stat], Mar. 2017. Accessed on: Feb. 27, 2021. [Online]. Available: http://arxiv.org/abs/1606.03490. 3

[13] S. Lundberg and S. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. Long Beach, California, USA. 3

[14] S. Singh M. T. Ribeiro and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. arXiv:1602.04938 [cs, stat], Ago. 2016. Accessed on: Feb. 27, 2021. [Online]. Available: http://arxiv.org/abs/1602.04938. 3

[15] C. Molnar. Interpretable machine learning. a guide for making black box models explainable. 2019. Accessed on: Feb. 27, 2021. [Online]. Available: https://christophm.github.io/interpretable-ml-book/. 3

[16] C. Padole and H. Proença. Periocular recognition: Analysis of performance degradation factors. *Proceedings of the Fifth IAPR/IEEE International Conference on Biometrics – ICB 2012*, 2012. New Delhi, India. 4

[17] H. Proença and J. C. Neves. Deep-prwis: Periocular recognition without the iris and sclera using deep learning frameworks. *IEEE Transactions on Information Forensics and Security*, 13(4):888–896, 2018. 3

[18] H. Su M. Bennamoun S. Minaee, A. Abdolrashidi and D. Zhang. Biometrics recognition using deep learning: A survey. arXiv:1912.00271 [cs], Fev. 2021. Accessed on: Feb. 27, 2021. [Online]. Available: http://arxiv.org/abs/1912.00271. 1

[19] B. C. Dhara R. K. Rout S. Umer, A. Sardar and H. M. Pandey. Person identification using fusion of iris and periocular deep features. *Neural Networks*, 122:407–419, 2020. 3

[20] B. Mittelstadt S. Wachter and L. Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99, 2017. 2

[21] S. Laine T. Karras and T. Aila. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*

(CVPR), pages 4396–4405, 2019. Long Beach, CA, USA. 4

[22] A. Ross U. Park and A. K. Jain. Periocular biometrics in the visible spectrum: A feasibility study. *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–6, 2009. 3

[23] 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*, 12(5):1017–1030, 2017. 3, 7, 8

[24] 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*, 13(12):2937–2952, 2018. 3, 7