Deep Adversarial Framework for Visually Interpretable Periocular Recognition

João Brito and Hugo Proença
University of Beira Interior, Portugal
joao.pedro.brito@ubi.pt, hugomcp@di.ubi.pt

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Sit amet nisl suscipit adipiscing. Eu volutpat odio facilisis mauris sit. Commodo sed egestas egestas fringilla phasellus faucibus scelerisque. Eleifend mi in nulla posuere sollicitudin aliquam. Sagittis orci a scelerisque purus semper eget duis at tellus. Amet est placerat in egestas erat imperdiet. Dui ut ornare lectus sit amet est placerat. Quam pellentesque nec nam aliquam sem et. At urna condimentum mattis pellentesque id nibh.

Index Terms – *Interpretability, Periocular Recognition*

Figure 1: Interpretable explanation provided by the proposed method. Areas highlighted in whiter tones provide reasons as to why the two samples come from different subjects (i.e. eyebrows, eyelashes and a skin spot). Underneath the visual answer, a text caption categorically classifies the pair in a traditional manner.

1. Introduction

This work focuses on the periocular recognition task, aided by easily interpretable explanations. In this context, one should consider the system's accuracy, as well as, its ability to portray the reasoning that supports a match/non-match decision. The latter is becoming an integral part of Machine Learning systems, given how ubiquitous and dependable they have become in recent years [13]. Thus, we diverge from the black-box paradigm and embrace an interpretable nature, as seen in figure 1.

In broad terms, a recognition task involves a set of unique and non-transferable features that, when given to a system designed to do so, can unmistakably identify a subject. Biometrics, as they are designated in the field, serve such purposes, as long as they are universal, sufficiently distinguishable, resilient to changes and realistically easy to collect [4]. Upon proving compliant with the aforementioned requirements, biometrics can be separated into two dominant categories. On one hand, *physiological* features like the irises, fingerprints and retinas are naturally possessed by a given subject. On the other hand, the gait and signature typify *behavioural* biometrics, due to the fact that they are manifested whenever a subject interacts with the surrounding environment [8].

As a subordinate field, periocular recognition makes use of the rich area comprising and surrounding the eye, in which the iris, sclera, eyebrow, eyelid and skin stand out with major relevance. We also acknowledge this set of biometrics, considering, where applicable, the colour and/or shape of said components.

Regarding interpretability and its applicability in recognition tasks, let us not forget 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 opaque algorithms gamble on the clearance of unauthorised people into sensible areas. It becomes clear, then, that including interpretability in AI systems is imperative. More recently, the politicians have addressed this urgency in formal terms. Namely, the EU, through the GDPR [2], introduced the notion of "right to an explanation". The definition and scope are still subject to debate [19], but these are definite strides to regulate the interpretable depth of autonomous systems.

Following the above paragraphs, this paper describes a framework that receives a pair of images (either *genuine* or *impostor*, depending on whether the images come from the same subject or not) and produces a twofold output: a binary match/non-match decision and a visually interpretable explanation (in this case, it takes the form of an image).

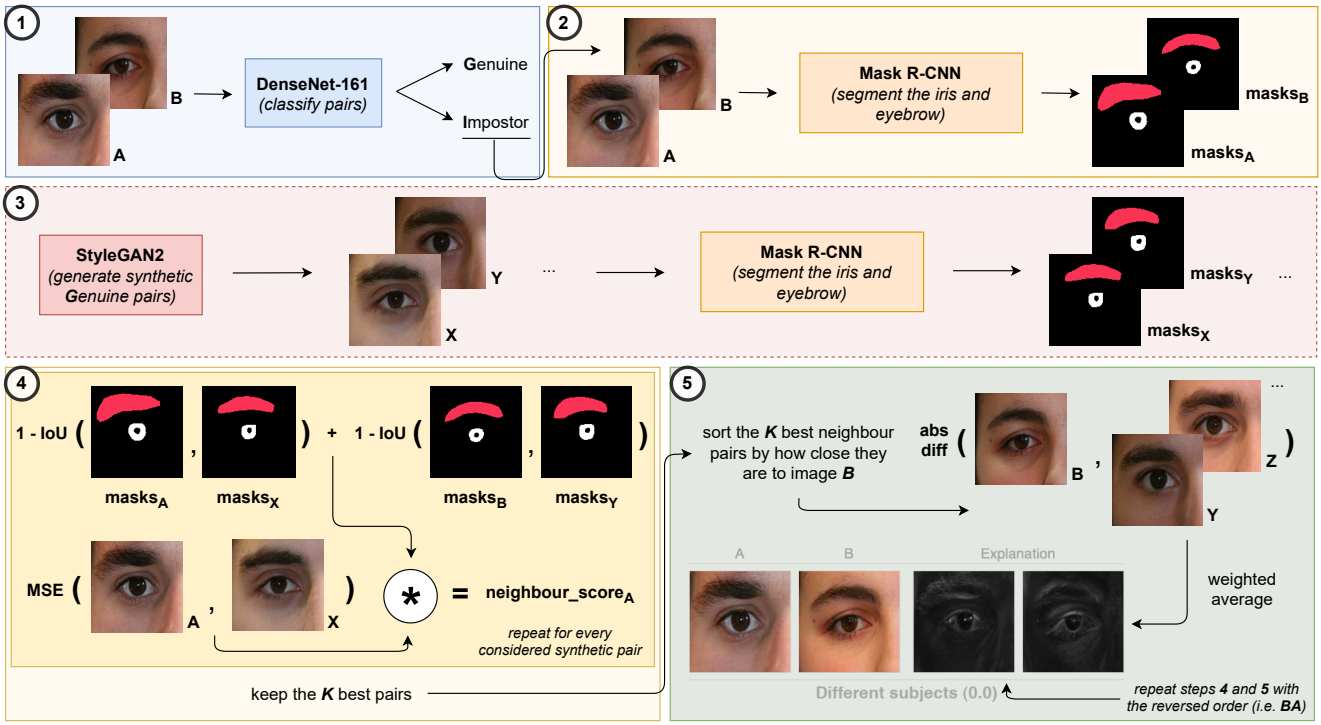

Figure 2: Diagram of the main pipeline. Step one encompasses a CNN that distinguishes between *genuine* and *impostor* pairs. The latter are explained with steps two to five, in which we try to find *genuine* synthetic pairs that closely resemble the test pair. By doing so, and despite looking similar, the test pair will probably contain certain internal differences (i.e. between images A and B) that the synthetic ones do not, thus providing an interpretable explanation.

This can be seen as the main contribution of the present work, in that it creates an accurate and interpretable solution. Other important insights include the use of the powerful generative capabilities that GANs possess to create samples that wouldn't otherwise exist in the training data, thus giving it more variety and flexibility (such process can be seen as a form of data augmentation).

Figure 2 illustrates the main steps that enable the periocular recognition task and posterior explanation: a CNN is trained to optimally produce a match/non-match decision and, if the pair is deemed to be *impostor*, the next major step searches, amongst the synthetic dataset that the GAN generated, for similar looking *genuine* pairs. The key here is that, if the test pair has significant disparities between images A and B (that make up the pair), the closest synthetic pairs do not (they are *genuine*). Then, by simply computing pixel-wise differences between the test pair and the closest neighbours, those disparities become evident. This assumption is justified by the CNN's non-match decision, which can likely be attributed to common differentiating factors, such as the iris colouration, presence/absence of skin spots or eyebrow thickness.

To fulfil the goals described so far, the subsequent sections are organised as follows: section 2 summarises the research efforts in the fields of periocular recognition and ML Interpretability; section 3 describes our method; section 4 analyses the main results and section 5 closes this paper with some final remarks.

2. Related Work

2.1. Periocular Recognition

The seminal breakthroughs in periocular recognition tasks can be traced to a set of methods termed *feature descriptors*. Methods such as HoG, LBP and SIFT were able to produce simplified representations by extracting useful information regarding edges, texture and keypoints, respectively. In [12], the results from each feature descriptor are fused to provide a more comprehensive description of an image's content. This work served as the basis for consequent fusion based approaches, as in [16]. Extending towards the field of supervised learning, in [10] a Restricted Boltzmann Machine is used to learn a probabilistic distribution over the input data, further discriminated with metric

learning and SVMs.

With the realistic applicability of Deep Learning schemes, researchers turned to popular architectures, like CNNs, in the pursuit of ever increasing recognition accuracy. Accordingly, in [21] the main concept involves the use of multiple CNNs that are trained to become specialised in classifying certain semantic information (e.g. gender, age, and more). Then, a score fusion process creates a unified architecture. In [14], the authors enforce a CNN to ignore the ocular region (due to this region's tendency to capture light reflections and other performance degradation factors) and rely more on the surrounding area (eyebrow, eyelid and skin). Going against the idea of only relying on either the ocular or periocular regions, in [18] the iris and periocular biometrics are separately explored for classification purposes, with the resulting scores being fused to reach a final decision. More recently, in [20] the authors bridge the gap between biometric recognition (in their case, facial) and interpretability, by learning feature specific filters that activate in a relatively consistent way, based on their preferred spatial location. However, a fusion of periocular recognition and interpretability remains largely unexplored, thus motivating the development of novel solutions.

2.2. Machine Learning Interpretability

In the literature, interpretable techniques are commonly discriminated in terms of depth, scope and model applicability [6, 9]. 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 interpretable (*intrinsic*) or allows complexity and focuses on explaining just the outputs (*post hoc*). Scope indicates the range that a technique possesses, i.e. if it explains individual predictions (*local*) or the model's entire behaviour (*global*). Finally, the applicability metric classifies techniques based on their model affinity, i.e. whether they are only compatible with a specific family of models (*model-specific*) or virtually any kind of model (*model-agnostic*). Commonly cited techniques include LIME [15] and SHAP [7]. 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 [17] 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 plotting reasons and, therefore, outside the preferred scope of this work, PDP [3] and ALE [1] are able to produce plots correlating independent variables and a dependent target variable, by exploiting the notions of marginal and conditional distributions, respectively.

3. Proposed Method

3.1. Method Description

The first step involves the use of a DenseNet-161 model to provide the binary part of the final answer: upon receiving a pair of images, the model's output is one of two classes (*genuine* or *impostor*). This step essentially performs the periocular recognition task. If the pair is deemed to be *impostor*, the remaining steps in figure 2 are responsible for creating the visually interpretable explanation.

According to the numbering system used, step two takes the test pair and, using an already trained Mask R-CNN model, segments the irises and eyebrows. Then, step three uses a trained StyleGAN2 generator to create a large, synthetic dataset of *genuine* pairs (i.e. where both images have differences but belong to the same person). For each of these synthetic pairs, a trained ResNet-18 model determines the side configuration (i.e. whether both images come from the left or right side of the face) and, as before, masks are obtained with the segmentation model. It should be noted that this step can be done beforehand and not during the inference stage. This is also the preferred behaviour, given how time consuming step three can be.

After obtaining the synthetic images and respective masks, the synthetic dataset is structured based on the iris positions, enabling faster search. To that end, the clustering algorithm K-Means is trained on a subset of the iris segmentation masks to compute three centroids, one for each major iris position (i.e. left, centre and right). With them, one can store the images based on their combination of iris positions (e.g. left-left, right-centre, etc...). By doing so, when searching, the algorithm can just use the synthetic pairs that share the same combination as the test pair, saving time and useless calculations. Obviously, the centroids provide a liberal classification of where the iris are. During the search process, the iris positions are subject to a stricter comparison.

Upon settling for a portion of the synthetic dataset that closely meets the iris position constraint, the segmentation masks can be further used to determine which generated pairs have the iris and eyebrow in approximately the same position as their counterparts in the test pair. Such precondition is key, given that pixel differences, which make up the visual explanation, are sensitive to component misalignment. With effect, a synthetic neighbour's score is given by the following formula:

$$score = w_{masks} * MSE(test\_pair_A, neighbour_X) \quad (1)$$

In equation 1, it becomes clear that a weighted distance is measured for each synthetic neighbour, with respect to the first images (i.e. image A for the test pair and image X for the synthetic one). Moreover, w_{masks} influences the computed MSE score to favour pairs that have good com-

ponent alignment and penalise those that do not (to achieve this, a factor equal to $1 - IoU(.,.)$ is used). Then, if the score is better than the currently saved pairs, the considered pair is kept as one of the best matches, up to that point. This iterative process continues until every synthetic pair is analysed. Such process tries to find, amongst the thousands of synthetic pairs, the ones that are closer to the test pair, in terms of the first images. Therefore, given that the second image of the test pair is not a *genuine* match, it will most likely be different in some areas to those of the optimal synthetic neighbours, and that is exactly the kind of dissimilarities that make up the final explanation.

From this point forward, the K closest neighbours are sorted by how close they are to image B, using equation 1 (computing the MSE score with respect to the second images, instead). Finally, to produce the final explanation, the K best neighbours are used to compute pixel differences against the test pair's image B. In practice, a neighbour's distance is subtracted from the total sum of distances, thus creating an inverted distance. This approach ensures that neighbours with smaller distances receive more weight as opposed to those with bigger distances. Then, the inverted distances are simply divided by the sum of inverted distances so as to normalise them. The final difference image results from those inverted distances acting as weights to determine the importance of each intermediate difference, resulting in an interpretable explanation that highlights crucial disparities.

3.2. Implementation Details

The DenseNet-161 model is trained for 15 epochs with a learning rate of 0.0002 and a batch size of 64 image pairs. The Adam optimiser is responsible for the weight optimisation process (with default β_1 and β_2 values). The CNN responsible for determining the side configuration of a given pair (i.e. ResNet-18) is trained with almost the same parameters (except for the number of epochs, which is equal to 5). The Mask R-CNN's training process uses almost all the default values, translating into 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). As for the StyleGAN2 architecture, its training comprises a total of 80000 iterations and a batch size of 8. The size of the generated synthetic dataset roughly settled on 400000 pairs, a number that, if increased, can very well improve the overall results (more samples enable more variance and possibly better neighbour matching). Finally, the number K, which determines how many synthetic neighbours should be kept, receives a default value of 15.

4. Experiments and Discussion

4.1. Working scenario

As is customary with biometric recognition systems, it is important to define the *working mode* and *world setting*, upon which the work is built. With respect to the *working mode*, a system is said to be in verification mode (also referred to as *one-to-one*) if it tries to validate a claimed identity (i.e. subjects identify themselves and the system's task is to validate those claims by comparing the extracted features to those stored in a database). On the other hand, if in identification mode (*one-to-many*), a system will try to retrieve the most likely match, within a know set of possible IDs [4]. As for the *world setting*, it discriminates between systems that can only accept a pre-defined set of IDs (closed-world) and those that continue to function even when new, unexpected subjects appear (open-world).

Based on the definitions above, our method essentially performs verification (the comparisons are done *one-to-one*) within an open-world setting, meaning that unseen subjects do not impose compromises or limitations.

4.2. Datasets

As mentioned above, the proposed framework consists of two modules, one for recognition and the other for interpretation purposes. Regarding the former, the chosen CNN is solely trained on the UBIPr dataset [11], which is naturally oriented towards periocular recognition problems and contains valuable ID information. As for the latter, it mainly relies on a combination of both UBIPr and FFHQ [5]. Despite not being immediately applicable to the context of this work (i.e. it contains full face images and, therefore, requires extra steps to extract just the periocular region), the FFHQ dataset contains unquestionable variety in terms of attributes, some of which are scarcer in the UBIPr dataset. In practice, a small, but curated, portion of the FFHQ samples is used to create a more varied super set, which favours the training of GANs. Regardless of their source, all images were resized to a common size, depending on the task (i.e. 512x512x3 for Mask R-CNN and 256x256x3 for StyleGAN2).

4.3. Qualitative Evaluation

Like most interpretable results, the proposed method produces images whose accuracy relies on some degree of subjectivity. Nonetheless, to create a comparison scheme and considering the lack of a direct SOTA comparison, a DenseNet-121 model, trained to perform the verification task, is paired with three commonly used interpretability techniques: LIME, SHAP and Saliency Maps.

Figure 3 displays the expected results from a visually interpretable system. LIME tries to keep the most important super-pixels, while SHAP highlights those it deems impor-

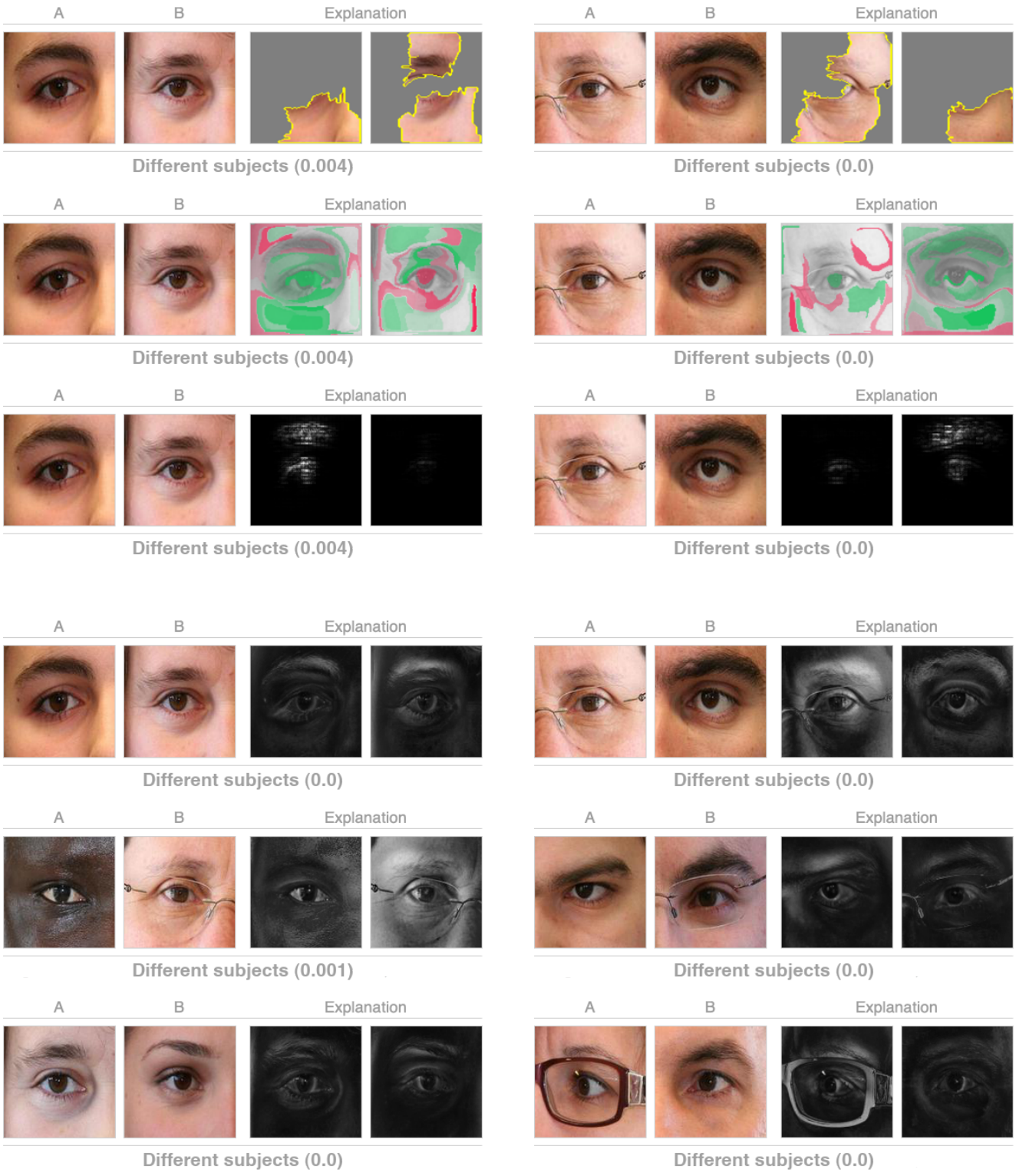

Figure 3: Interpretable results from LIME, SHAP, Saliency Maps and ours. The top three rows contain the three interpretability techniques, while the bottom three display our results (with two directly comparable pairs and four additional ones). Notice how the eyebrows and skin are accurately highlighted.

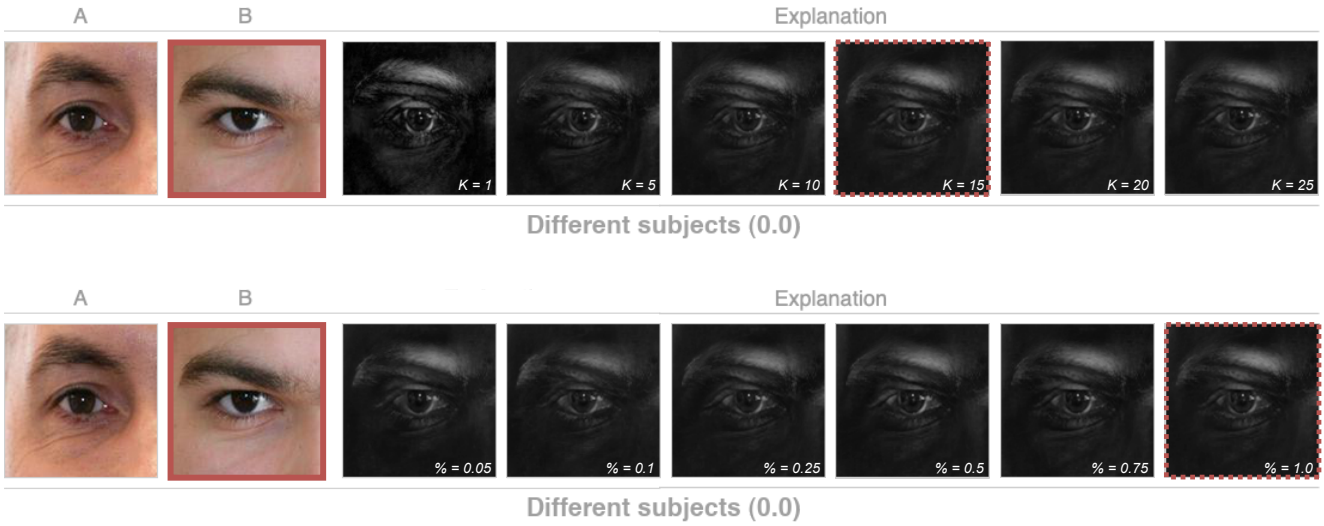

Figure 4: Visual effects when key hiper-parameters are changed. As highlighted in red, image B is the one being explained. Additionally, red dashed squares depict the default values. In general, increasing K allows for smoother explanations, while decreasing the size of the generated dataset does the opposite.

tant in green tones. Saliency Maps and our method try to produce greyscale explanations. In general, our results are the most informative, when compared with the remaining three solutions. Focusing more on what is achievable with the proposed method, the first sample typifies the desired level of interpretability: the eyebrows are different in terms of thickness, there is a skin spot on just one of the subjects and the eyelashes are also positively distinct. Moreover, the third sample (right below the first) clearly highlights the entirety of the skin area, which is obviously different between images A and B. Finally, the fourth sample also shows how the eyelids differ, by colouring subject B's eyelid.

4.4. Quantitative Evaluation

TODO

4.5. Ablation Study

As mentioned before, the two major hiper-parameters of the proposed method are K and the length of the synthetic dataset. Changes to these values can affect the quality of the generated explanations in a less than optimal way (as seen in figure 4).

4.5.1 Number of closest neighbours

The value K determines how many synthetic pairs should be considered as the closest *genuine* neighbours to a given test pair. Smaller values lead to more sensitive and jagged results, unlike those achievable with larger values. Up to a certain point (e.g. 15), increasing K creates smoother explanations, due to the larger number of samples taken into account when averaging the intermediate differences. This trend, however, starts returning incremental improvements that do not justify themselves.

4.5.2 Synthetic dataset length

Considering the main goal of the search process is to find *genuine* pairs that closely resemble the given pair, restricting the amount of possible matches can impose certain limitations. More samples allow for better component matching, especially around the iris. Additional benefits also include a smoothing effect.

5. Conclusions and Further Work

References

[1] D. W. Apley and J. Zhu. Visualizing the effects of predictor variables in black box supervised learning models, 2019.

[2] E. Commission. 2018 Reform of EU Data Protection Rules.

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

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

[5] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks, 2019.

[6] Z. C. Lipton. The mythos of model interpretability, 2017.

[7] S. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions, 2017.

[8] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang. Biometrics recognition using deep learning: A survey, 2021.

[9] C. Molnar. *Interpretable Machine Learning*. 2019. Available from: `https://christophm.github.io/interpretable-ml-book/`.

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

[11] C. Padole and H. Proença. Periocular recognition: Analysis of performance degradation factors in proceedings of the fifth iapr/ieee international conference on biometrics – icb 2012, new delhi, india. 03 2012.

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

[13] A. Preece, D. Harborne, D. Braines, R. Tomsett, and S. Chakraborty. Stakeholders in explainable ai, 2018.

[14] 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.

[15] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.

[16] A. Ross, R. Jillela, J. M. Smereka, V. N. Boddeti, B. V. K. V. Kumar, R. Barnard, X. Hu, P. Pauca, and R. Plemmons. Matching highly non-ideal ocular images: An information fusion approach. In *2012 5th IAPR International Conference on Biometrics (ICB)*, pages 446–453, 2012.

[17] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2013.

[18] S. Umer, A. Sardar, B. Dhara, R. Rout, and H. Pandey. Person identification using fusion of iris and periocular deep features. *Neural Networks*, 122, 11 2019.

[19] S. Wachter, B. Mittelstadt, 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, 06 2017.

[20] B. Yin, L. Tran, H. Li, X. Shen, and X. Liu. Towards interpretable face recognition, 2019.

[21] 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.