Image and Vision Computing 132 (2023) 104645

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Visual and textual explainability for a biometric verification system based on piecewise facial attribute analysis

Lucia Cascone [a,*], Chiara Pero [a], Hugo Proença [b]

[a] Department of Computer Science, University of Salerno, Italy
[b] Department of Computer Science, University of Beira Interior, IT: Instituto de Telecomunicações, Portugal

ARTICLE INFO

Article history:
Received 8 September 2022
Received in revised form 13 December 2022
Accepted 25 January 2023
Available online 05 February 2023

Keywords:
Interpretable representation
Explainability
Feature extraction
Semantics
Facial attribute analysis

ABSTRACT

The decisions behind the mechanics of a biometric verification system based on Machine Learning (ML) are difficult to comprehend. Although there is now well-established research in various fields of application, such as health or justice, the use of ML-based methods is accompanied by a lack of confidence that results in their limited use. The explainability of a ML system and the comprehension of what lies behind its prediction is one of the numerous characteristics that define "trust" in these systems. Over the years, face-based biometric authentication has been the subject of extensive research in both academia and industry. However, existing biometric authentication systems still have problems regarding accuracy, robustness and, explainability. Still lacking in the literature is a comprehensive examination of the use of post-hoc explainability techniques for such systems. Cognitive neuroscience has always been interested in the method by which people perceive faces; local elements such as the nose, eyes, and mouth are critical to the perception and recognition of a face. In this work, starting from this assumption, we propose a framework of visual and textual explainability based on the parts of a face by analyzing them with respect to the facial attributes reported in the CelebA dataset. The primary objective is to be able to explain why two pictures of different subjects are distinct. This is done by sinthesizing pairs of images that illustrate how dissimilar the various parts of the face under investigation are and incisive and direct textual explanations of the distinguishing features are generated. A further study analyzes an interpretable mapping between the semantic space of the text and the space of the image.

1. Introduction

In recent years, Machine Learning (ML) algorithms have been widely utilized in several research fields, including healthcare, and even as support in the administration of public order. Even though the benefits of using these kinds of systems have been well documented in the literature and from a theoretical point of view, their use in the real world is still behind. This is also due to the lack of transparency that frequently accompanies the development and implementation of such innovative technologies, which the populace views with skepticism and disillusionment. In fact, these models are often more accurate, but they work like impenetrable black boxes. This makes it hard to understand the basic ideas behind their predictions. Instead of trying to make models that are naturally easy to understand, there has been a lot of recent work on "Explainable ML," which uses different methods to explain why a model made a certain decision [1]. Facial feature analysis can be utilized as supplemental (or soft) information to enhance the performance of conventional biometric systems and assist those that generate a textual description for facial recognition, hence reducing the identification search space. By verbalizing the description based on facial characteristics, it is possible to give the user a tool that simulates human conversation when attempting to identify or differentiate a person. Thus, the purpose of such a model is to facilitate communication that more closely resembles human communication by emphasizing distinctions through textual descriptions. As a rule, humans prefer concise explanations that compare the current situation to one in which the event did not occur. Especially uncommon causes provide adequate explanations [2]. In this research, we intend to propose a post-hoc framework for the explainability of a biometric identity verification system. To compare the two identities, both a verbal description based on the dissimilarities between the facial traits and an image illustrating the differences between the 21 characteristics of each identity are provided. Additional study reveals the relationship between the textual space and the image space, as well as the potential of the two systems in relation to the limitations of the dataset.

The rest of the paper is organized as follows: research on facial attributes and ML explainability is reviewed in Section 2. Section 3 presents

* Corresponding author.
E-mail addresses: lcascone@unisa.it (L. Cascone), cpero@unisa.it (C. Pero).

https://doi.org/10.1016/j.imavis.2023.104645
0262-8856/

the dataset in detail, while Section 4 provides a description of the method's most significant components. Finally, Section 5 analyzes the results obtained, and Section 6 draws the conclusion of this work and addresses open issues.

1.1. Motivations

Many ML models are "black boxes" that do not explain their predictions in a way that humans can comprehend, resulting in a lack of confidence in their potential use in the healthcare and justice fields. Therefore, an explainability architecture for a biometric identity verification system could be a tool for increasing human confidence in the system, as it would provide insight into why certain decisions were made. There are still few works in the literature that aim to provide an explanation in the field of biometrics. To deliver a relevant and valuable explanation, the model must provide an explanation that is both comprehensible and simulates what a human subject would emphasize. In this work, we focus on the analysis of facial characteristics and the intrinsic and extrinsic information that allows us to distinguish between two individuals. It is common knowledge that the information acquired from the face helps individuals to recognize the identity of the other, understand what he or she is feeling and thinking, forecast their actions, identify their emotions, develop relationships, and communicate through facial movements. Consequently, the purpose of this research is to propose, for the first time, an explainability framework for a biometric verification system based on facial characteristics.

1.2. Main contributions

This paper proposes a post-hoc explainability approach to assist users in understanding decisions made by a biometric verification system that uses information from individual parts of the face to determine the mismatch between two different identities. In fact, the goal of the framework is to give meaning to the decisions made by a system based on the main attributes and features of the face, thus providing an explanation as to why the images of two different subjects do not match. The user will be able to have both an image and a textual caption that makes explicit the two main attributes or characteristics that differentiate two different subjects, the so-called impostors, and will be able to have an interpretable mapping between the semantic space of the text and the space of the image. The image makes it possible to visualize the distance between the 21 facial attributes being studied and objects in a two-dimensional space.

2. Related works

2.1. Facial attribute analysis

The human-understandable visual characteristics of face images are described by facial attributes, which represent intuitive semantic features [3]. Facial attribute analysis is widely used in real-world applications such as face verification, face identification, face retrieval, and face image synthesis. It is well known that some features of the face, such as soft biometric traits, can enhance the functionality of traditional biometric systems and facilitate human-explained recognition. With the effective application of Deep Learning solutions, researchers turned to popular architectures to pursue ever increasing performance.

Kumar et al. [4] introduce one of the first facial attribute analysis methods for face verification, developing two approaches using traits computed on face images based on describable attributes and simile classifiers. Next, the authors in [5] extracted a diverse set of highly discriminative intermediate features, namely Part-based One-vs.-One Features (POOFs), followed by linear SVMs for each attribute. Preliminary results in using Deep Learning for face verification and predicting facial attributes were obtained, respectively, by Chung et al. [6] and Liu et al. [7]. In [8], the authors proposed a new Deep Learning framework for learning facial attributes by exploiting videos and contextual data captured by a wearable sensor. Zhong et al. [9] adopted the mid-level CNN features for face attribute prediction, based on the observation that facial attribute characteristics are different: some of them are locally oriented while others are globally defined. The task of facial attribute prediction has also been investigated as a regression problem with a 16 layer VGG topology in order to minimize the mean squared error loss [10]. Abate et al. [11] described an unsupervised clustering approach for face attribute recognition. The proposed method is a neural network model based on transfer learning. The goal of this model is to group faces based on the facial features they have in common. A novel Deep Learning formulation for facial attribute analysis, called R-Codean autoencoder, was presented in [12]. Recently, several deep CNNs architectures have been developed for performing multi-attribute classification [13–15].

2.2. Explainability in machine learning

The goal of ML interpretability is to close the gap between a system's predictions (i.e., the what) and the rationale behind those predictions (i.e., the why) [16,17]. According to [18], the following criteria are used to categorize explainable approaches: depth, scope, and model applicability. The degree of complexity of a model is represented by its depth. Contrary to post-hoc approaches, which allow complexity and simply seek to explain the model outputs, intrinsic techniques frequently place limits on a model's complexity. Scope represents the range of an interpretability approach. A technique is working on a local scale when it can explain specific predictions. On the other hand, a strategy performs a global explanation if it enables us to comprehend a model at once. Finally, a technique's capacity to explain families of models or architectures is referred to as model applicability. Because they depend on specific traits of a given type of model, model-specific techniques are only applicable to that type of model. On the other hand, model-agnostic methods are so general that they can be used with almost any model. Furthermore, these methods can be differentiated using various but equally valid criteria. It's crucial to assess an interpretability approach according to its expressiveness and complexity. While the latter focuses on the computing cost of generating explanations and may make some techniques inapplicable, the former is bound to the area in which the explanations exist (for example, natural language or images). Strategies can also be set apart by the kinds of explanations they can give. The three criteria that stand out in this context are stability (i.e., differences between explanations of slightly different samples), accuracy (i.e., whether the explanations are correct for unobserved data), and comprehensibility (i.e., the level of difficulty in trying to interpret an explanation). In literature, there are various forms of interpretability that are all equally valid [19]. Plotting the correlations between an independent variable and a dependent variable is a common practice for techniques like PDP [20] and ALE [21], whereas LIME [22] and SHAP [23] provide images with highlights in certain super-pixels (i.e., groups of neighboring pixels).

3. Large-scale CelebFaces attributes (CelebA) dataset

The experiments are conducted on CelebFaces Attributes (CelebA) Dataset [24], which has roughly 200 k celebrity images with a total of 10.177 identities, each with about 20 frames. CelebA is a large-scale, fully annotated database of facial attributes that has been frequently used in literature for predicting facial attributes, detecting faces, and locating landmarks. Every face image is annotated with 40 binary attributes and 5 key points to align the image to 55×47 pixels. The wide range of environmental and behavioral factors such as pose, expression, ethnicity, age and gender, and occlusion variations, make CelebA a very demanding dataset. Furthermore, it is possible to observe how the attributes are highly imbalanced with respect to each other. In detail, a third of the characteristics are very unusual facial traits (10% frequency or

less), with only a pair of them being exceptionally prevalent (present in over 70% of cases) [25]. As a result, the largest imbalance ratio between minority and majority facial attributes is 1:43. Another issue is characterized by the incorrect/noisy annotations; one can observe that there is a strong bias in favor of one of the two classes for many of the facial attributes. For example, this is particularly true for some characteristics, such as how the bulk of face images are categorized as "Young" and just a small number as "Bald" or "Wearing Hat" [26]. Sample images from CelebA dataset are shown in Fig. 1.

Our research focused on facial parts related to the periocular area (including eyes and eyebrows) the nose, mouth, cheeks, hair, chin, forehead, and the area upper lips. For each area, in the dataset annotation list, we identified several associated facial attributes that are shown in Table 1.

Table 1
Facial parts and corresponding attributes considered in our work.

Facial Part	Facial Attributes
Eye	Bags under eyes, Narrow eyes, Eyeglasses, Heavy makeup
Eyebrow	Arched eyebrows, Bushy eyebrows
Nose	Big nose, Pointy nose
Mouth	Big lips
Upper Lip area	Mustache
Cheek	5 o'clock shadow, Pale skin, Rosy cheeks
Chin	Double chin, Goatee
Forehead	Receding hairline
Hair	Black hair, Blond hair, Brown hair, Gray hair, Bald, Bangs, Wavy hair, Straight hair, Sideburns

4. Proposed method

4.1. Semantic exploration of facial attributes

4.1.1. Image segmentation

Image segmentation is the process of recognizing each pixel in an image as belonging to a specific category. Semantic segmentation is the difficult process of grouping pixels with the same label, i.e., those that share specified qualities, in computer vision. Face segmentation is the classification of a person's face into different categories, such as eyes, hair, nose, and lips, each of which can be treated separately. Face segmentation is a critical issue in face image analysis since it is required not only for understanding facial features but also for post-processing tasks such as virtual face make-up and virtual face swapping.

In this study, we use two different algorithms to separate the parts of the face we're interested in based on the labels for facial features given in the dataset. The first strategy is based on Deep Learning and also exploits the CelebAMask-HQ dataset [27], while the second one implements a geometric approach. CelebAMask-HQ is a collection of 30.000 high-resolution facial images chosen from the CelebA dataset. Based on the information in CelebA, a face attribute segmentation mask was associated with each image. The masks obtained (size of 512×512) were manually annotated on the basis of 19 classes, which included facial components such as skin, nose, eyes, eyebrows, and ears, but also accessories such as glasses, earrings, necklaces, hats, etc. In this paper, we use Mask R-CNN, a model pre-trained on that dataset, to extract for all CelebA images, masks related specifically to hair, nose, glasses, upper and lower lip, left and right eyebrow, and eyes. In CelebA, however, there is also information about the characteristics of some facial features (like whether or not a person has a mustache) for which there is no segmentation mask in the dataset. In order to generate them, a landmark based approach is also implemented. Using the Dlib-ml package [28], we produced 68 annotation points as the image's real landmarks.

For the area around the upper lip, the cheeks, and the forehead, new masks were generated by using previously created masks as well as the most pertinent facial landmarks to outline the missing parts.

After obtaining the following masks, those pertaining to the mouth (upper and lower lip) and the periocular area (right and left eye) were combined to gather a higher amount of information. Then, as a morphological level reconstruction, dilation was used to gradually broaden the borders of the mask regions in order to include any extra information lost during their definition. The masks were then overlaid on the original images to extract only the related regions, as shown in Fig. 2. After this operation, 9 images containing the individual facial components listed in the 1 table were generated for each image.

4.1.2. Latent space visualization

To get data from the facial parts of interest, an image was changed from its original space to its latent space. The idea of "latent space" is significant since Deep Learning relies on its application. Latent space corresponds to an abstract multidimensional space that encodes a meaningful underlying structure of the input data, representing it with peculiar values that we cannot evaluate directly. Autoencoders have become a viable option for doing this encoding.

Autoencoders seem to have minimal research on choosing the right number of latent space dimensions (k), a parameter that defines the input's quantitative spatial representation. In this paper, we fix such $k = 4092$ after several empirical assessments. More in detail, starting from the minimum latent space dimension adopted in literature, we try different k sizes that lead to maximum input reconstruction capacity and maximum utility for classification tasks. Even small changes to the latent space could lead to big changes in the original space of the observed data. Because the original space is much bigger to study, the model

Fig. 1. Sample images from CelebA dataset.

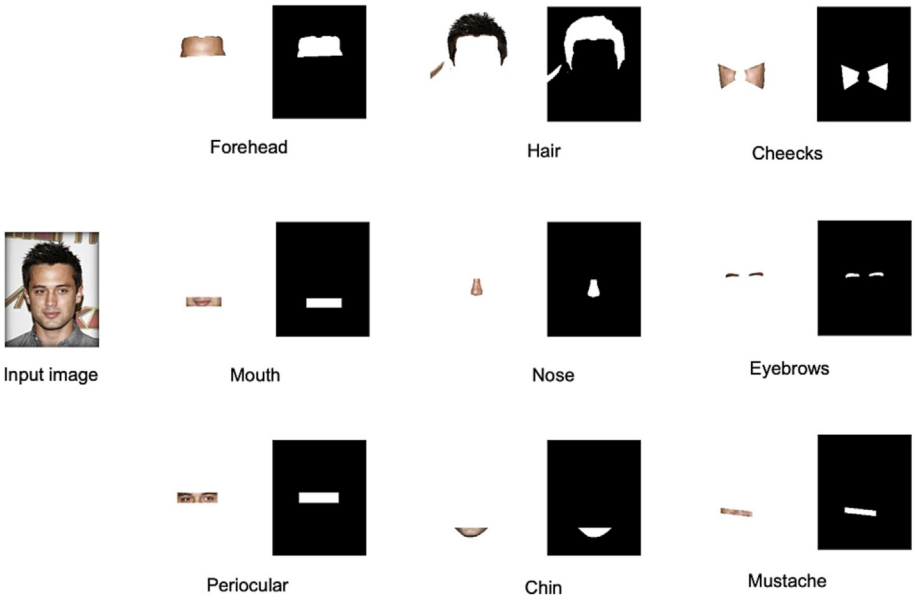

Fig. 2. Segmentation of a facial image into its semantically coherent regions where analysis is carried out.

would be better able to make sense of the observed data if it looked at latent space instead of the observed data itself. Autoencoders are in themselves a powerful dimensionality reduction algorithm. As a result, integrating them with the UMAP technique may improve the acquisition of latent structures even more. Uniform Manifold Approximation and Projection (UMAP) [29] is a popular nonlinear dimensionality reduction approach. It compresses data into a lower-dimensional space while retaining as much of the data's local and global structure as possible, while also reducing computation time. Our goal is to leverage this combination to have an interpretable visualization method for an initial exploratory analysis of the facial attributes under study. As a result, we use the UMAP technique to model the latent space data in a 2D representation. In contrast to the pixel space, the images in the latent space are shown in a simpler way. Using UMAP, it is possible to search for an interesting two-dimensional projection of latent space. This makes it easy to see and understand the structure of the data and any closeness in space between the data of subjects with the same attribute. Indeed, UMAP maintains the data density, neighborhoods, and distances. Therefore, for each subject, once the parts of the face of interest were extracted thanks to the constructed masks, the relative two-dimensional UMAP coordinates were generated for each of the 21 attributes using the vector of 4092 latent space components as input. For example, consider wishing to graphically depict the difference between different attributes specified for hair color (brown, black, blond, gray, none). We first apply the mask that allows us to isolate only this feature, then use the resulting image, which then shows only the hair, as input for the autoencoder model. The hair is then represented in latent space as a vector of dimension 4092. The UMAP algorithm then receives this vector as input and projects it into 2D space.

4.1.3. Genuine-impostor score

To acquire evidence that even after this reduction operation, the data continue to maintain and preserve what is the salient and distinctive information of each characteristic, we conducted a further study. The purpose was to investigate if the information taken from the attributes and then drastically downsized were still able to discriminate between two identities. For each of the 21 attributes, coordinate pairs are generated using UMAP. These were then concatenated to obtain a single vector of dimension 42. Each subject is then described by this vector. Next, several random image pairs were defined. The pair with different images of the same subject was labeled as “genuine”, while the pair consisting of images of different subjects was labeled as “impostor”. The number of genuine and impostor image pairs was balanced. From the 9341 identities, only those for which at least 4 images were present in the dataset were selected. The two feature vectors associated with one image of the pair, respectively, were concatenated and used as input for a binary classification system.

Random Forest classifier was used as the model. There are various advantages of employing a tree learning algorithm: training on large datasets with resistance to redundant variables or high correlation variables, which can cause overfitting in other learning algorithms. It is well known that, in general, if fully developed decision trees are used, one may run into the problem of overfitting because this type of algorithm does not generalize well to invisible data. With Random Forest we address this problem because the idea behind it is to use a pool of decision trees where the values in the tree are an independent, random sample. With a random split of the data where 70% is used for the training phase and 30% for testing, an 89% accuracy rate was obtained. Further studies were conducted by increasing the projection size using UMAP. It is worth noting that with several components equal to four, the accuracy is 90%, whereas with eight components, the accuracy is 91%. Doubling the number of components again did not result in a performance improvement, an indication that UMAP still allows the preservation of salient information even when operating with such a drastic reduction in the number of components. Fig. 3 provides the overall methodology adopted to obtain the genuine/impostor score. The corresponding ROC curves are shown in Fig. 4.

4.2. Automatic generation of image captions

CNN's ability to pull out features and an LSTM's ability to make text were combined to make the automatic image captioning. [30]. The basic idea is to have a system that takes a pair of images as input and gives textual descriptions that, in principle, emphasize the different facial components. To ensure optimal size and storage of the image pairs, this method needs some preliminary steps. The image pairs are downsized to the ResNet model's predetermined size (i.e., 224×224 pixels). The pictures and descriptions are then properly matched in the training step.

4.2.1. Learning phase

The two main parts of this architecture can be thought of as an encoder (the CNN) and a decoder (i.e., the LSTM). The first attempts to create a compact representation of the received images, while the second

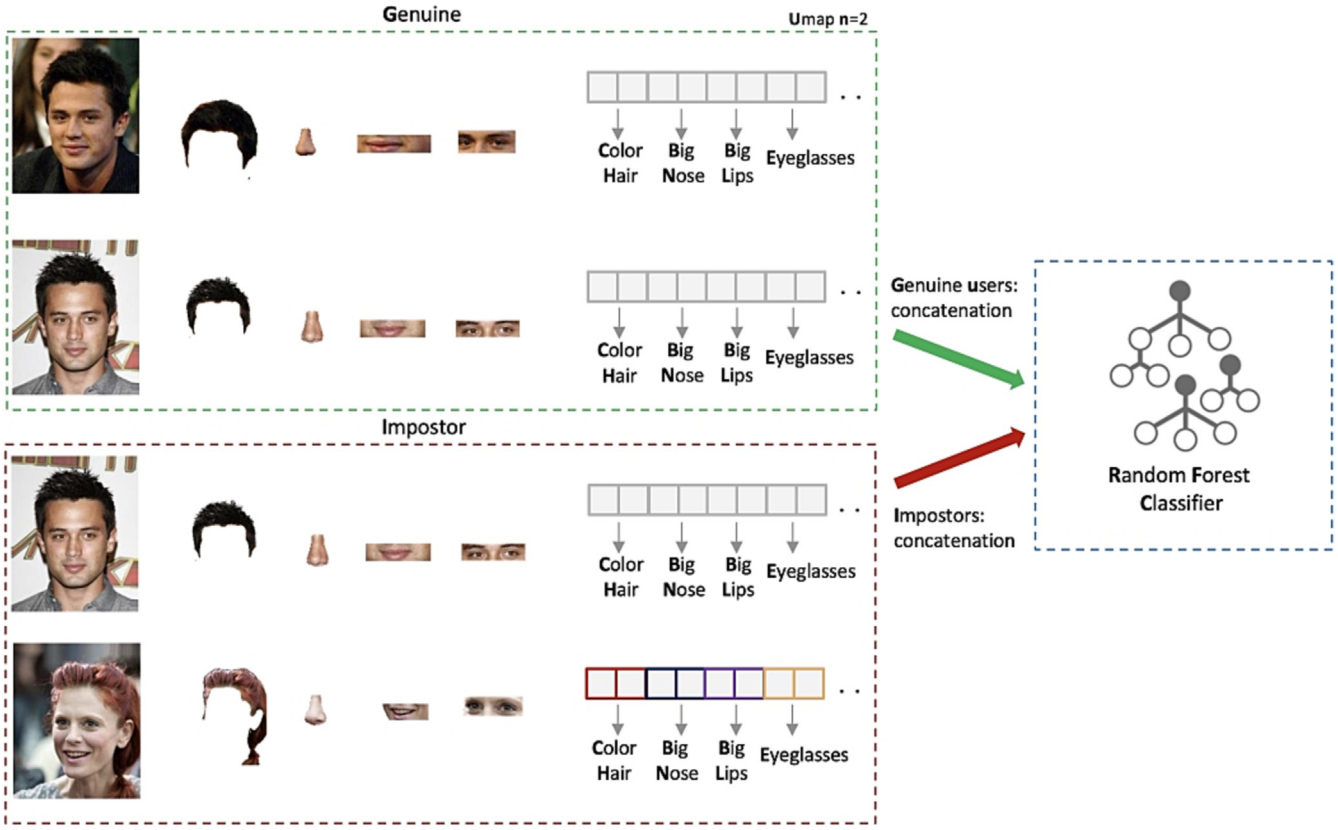

Fig. 3. Workflow of Genuine-Impostor. From left to right: creation of subject vector with 2D UMAP components extracted from facial masks, for a total of 21 attributes; creation of "Genuine" and "Impostors" image pairs; Random Forest classifier.

tries to provide an acceptable caption for that depiction. ResNet's feature extraction capabilities are used for the encoding. To make a single vector with 4096 values, the 2048-dimensional feature maps from both images were first added together. Then, a couple of linear layers were used to turn this into a 512-dimensional representation. The LSTM is responsible for predicting the tokens that would constitute a realistic caption, so the generated tensor is passed to it. The vector is then concatenated with an embedding of the ground-truth caption and fed via a padding operation, which reorders the input to get the desired shape.

4.2.2. Inference phase

The encoding stage is almost identical to the training stage: the same CNN is used for both input images. The feature maps are combined, and the feature vector is predicted by a couple of linear layers. Differences begin to surface during the decoding step. To begin, the LSTM takes only the feature vector and, using two linear layers, predicts the first token (ideally, "<START>"). The LSTM is given an embedding that keeps the token generation process going until one of the following stop criteria is met: 1) the expected sequence is longer than the predetermined maximum length, or 2) the token "<END>"is reached.

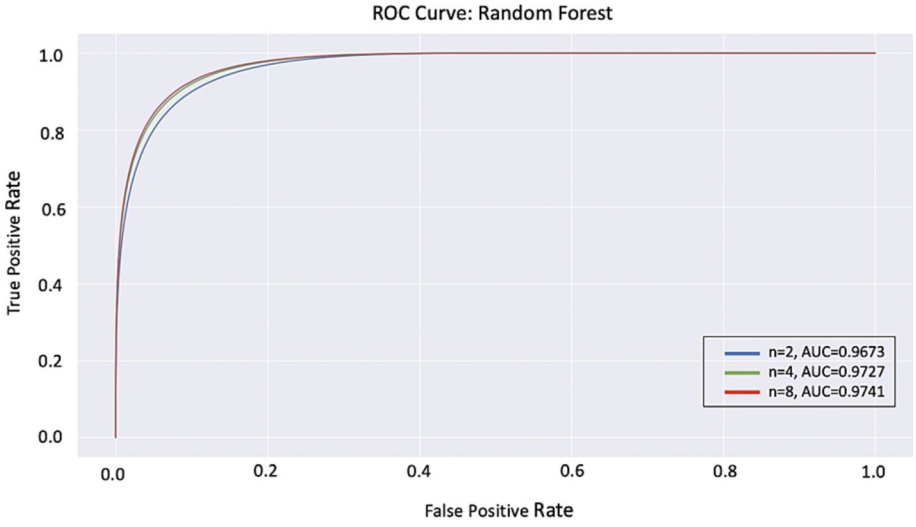

Fig. 4. ROC curve. *n* represents the UMAP projection size.

4.3. Implementation details

With a learning rate of 0.0001 and a batch size of 64 samples, the proposed method was trained for a total of 100 epochs. There were around 1855 unique pairs of training examples available, each with 4 captions. The Adam optimiser was used to update the weights; and we fine-tuned the CNN architecture weights from the ImageNet dataset [31]. Finally, the embedding length and hidden size of the LSTM were both set to 512.

5. Explainability discussion

5.1. Explainability evaluation

Most of the existing authentication methods based on Machine or Deep Learning can be seen as a black box, that is, they do not provide an explanation or justification for the obtained results. This significantly decreases the transparency of the whole system, and the user is limited to accepting its decision without clearly understanding why. In the method proposed in this paper, we refer to a post-hoc explainability approach to help the user obtain an explanation for decisions made by a system that exploits information from individual parts of a face to determine the inconsistency between two different identities. Indeed, the goal is to provide an interpretation to a system that is designed to highlight the main facial attributes/features that explain the decision of mismatch between two different subjects' images. Not only will the user have the opportunity to have a caption that makes explicit the two main attributes/features that differentiate two different subjects, but it will also be possible for him to establish an interpretable mapping between the semantic space of the text and the space of the image. Biometric recognition occurs based on individual parts of the face. In fact, once these parts have been extracted (as described in Section 3), they are each processed by an autoencoder, which gives a raw representation in latent space. In this way, it is possible to capture only the most representative information from the input, effectively ignoring nonessential or noisy information such as outliers. UMAP provides meaning to retrieved vectors so humans may interpret them. This is a dimensional reduction technique that provides a low-dimensional representation of the feature vector obtained from the autoencoder while still allowing the global structures of the original feature space to be maintained. Through a two-dimensional representation, it is then possible to project the facial attributes/features into space and visualize them through a point facilitating human comprehension with a direct and intuitive explanation. The projection in the two-dimensional space of the latent vectors related to the same attributes tends to minimize the distance between them. Thus, having taken the input pair of images, the face parts related to the features shown in Fig. 2 are first extracted for each. These are then used as input to an autoencoder algorithm to obtain their representation in latent space. Using the UMAP method, for each of the 21 attributes listed above, it is possible to get a projection of their representation in latent space into two-dimensional space. The goal is to reflect the global structure of the data as accurately as possible, tending to group similar categories together. So, for each image, 21 pairs of coordinates are found. Each pair shows how one of the attributes is projected into a two-dimensional space. By considering the lowest Euclidean distance between the related pairs, it is possible to evaluate what characteristics the two subjects share and where they differ most. In Fig. 5 we show an example in which, concerning the pair of images taken as input, we report a graph ordered concerning the value of the Euclidean metric. It can be observed that the distance between the two-dimensional projections of the relative vectors in the latent space is the greatest for hair color and whether or not a person has a mustache. The implemented text generator system automatically adds a caption to the image that shows which of these two characteristics the model thinks is best at showing how the two subjects are different.

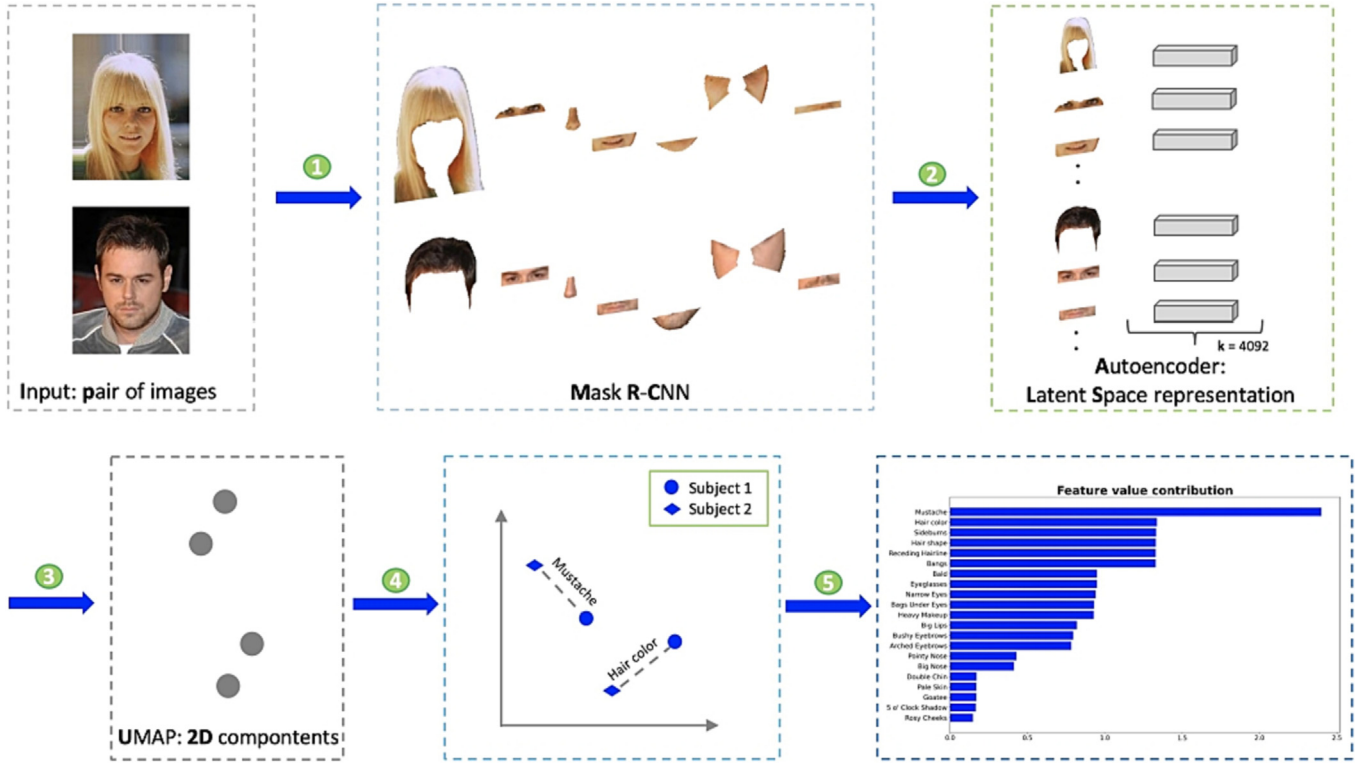

Fig. 5. Workflow: left to right, top to bottom. 1) Pair of images of two different subjects given as input 2) Extraction of face parts through the Mask R-CNN model 3) Use of an autoencoder for each previously extracted face part to obtain a projection of it in latent space 4) Projection through the UMAP algorithm in two-dimensional space of the representative vectors in latent space, with respect to the facial attributes under analysis 5) Calculation of the Euclidean distance between the coordinate pairs of the two subjects representative of the same attributes 6) Ordered graph showing the attributes for which the two subjects differ the most.

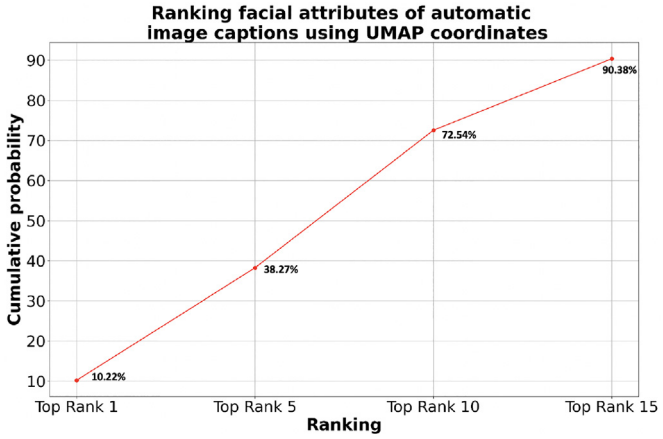

Fig. 6. Test set evaluation: correlation between the face components mentioned in the explanations and the euclidean distances between the UMAP 2D coordinates.

Thus, in Fig. 6 we can see how there is a strong correspondence between this visual relationship and the attributes/characteristics that are produced to provide a verbal explanation for the mismatch decision between the so-called "impostors." As a validation of this approach, we can indeed observe that for more than 10% the two features that are predicted by the text generator system are the ones that are the most distant in the graphical visualization using UMAP, and almost 40% are among the top 5 most distant, more than 70% when considering the 10 most distant up to more than 90% when considering 15. It should also be noted that these percentages include the pairs of images for which it is not enough for only one of the predicted attributes to be in the top k positions according to the Euclidean metric. This must be true for both attributes in the caption.

For example, for top rank *1*, if one of the two features predicted in the text generator had the third highest distance, this sample was not taken into account in the percentage.

A further analysis is depicted in Fig. 7. Indeed, it shows that the distribution of features in the training dataset that, as mentioned earlier, were extracted and validated by a human subject, and then actually predicted, is quite robust. From a cross-comparison, it was observed that the feature mainly used to detect imposters with both methods is the lips.

5.2. Success and failure case analysis

To conduct an in-depth performance analysis and critical interpretation of the results, the objective of this section is to gain an understanding of both successful and unsuccessful scenarios. The success and failure cases are studied both from the point of view of textual explanation and from the point of view of the ordered graph showing the differences between the two subjects. Fig. 8 shows four different pairs of images, each of which allows us to illustrate a different potential scenario that we faced. The first pair, pair 14, falls outside of the top ten ranks. We observe in the histogram that the characteristic deemed most peculiar is "big lips," which, when examining the annotated images in the dataset, gives ample opportunity for interpretation. The second, on the other hand, is "double chin," which is not detectable in the second image due to the obvious occlusion of the area. The encoded information of the mustache is strikingly similar for both images, which is in line with the image we are looking at. The prediction of the characteristics by the text generator, on the other hand, appears to be both consistent with respect to the distribution of attributes in the training dataset (Fig. 7) and discriminative for the two images, based on a visual check. In contrast, the second pair, 59, compares two images of two distinct women. When comparing the textual prediction with the information derived from the histogram, it is evident that makeup, which was supposed to be the most discriminating characteristic among those under examination, is also predicted by the textual model. On the contrary, the second prediction, relating to hair color, is not among the most discriminatory ones in the histogram. Examining the pictures in question makes it even more apparent that some of the problems revealed by the model are attributable to the improper labelling of some images; in fact, the woman in the second picture has red hair, which isn't in the list of facial attributes (see Table 1). The presence of bags under the eyes, the second most discriminating feature of the histogram, effectively distinguishes the two participants based on visual inspection. The third image pair (pair 156), which further highlights some possible issues related to the dataset, such as may be the presence of black and white images. So, when this happens, the two models are fooled, and the different color considerations should not be taken as always true. However, under particular lighting conditions, even some information related to shape geometry may be corrupted and not clearly visible as in this case for lip size. In the last image (pair 1) we instead record a success case, where both the characteristics predicted by the textual generator and the first two most discriminating attributes in the histogram coincide. The noticeable difference in hair color between the two persons, with the male having brown hair and the female having blond hair, and the presence or absence of a mustache are, also based on a visual examination, two quite prominent characteristics. From the study performed, it is clear that the results obtained in Fig. 6 may be more consistent with our proposed model and the text generator if a dataset with more accurate labelling and fewer wild images existed.

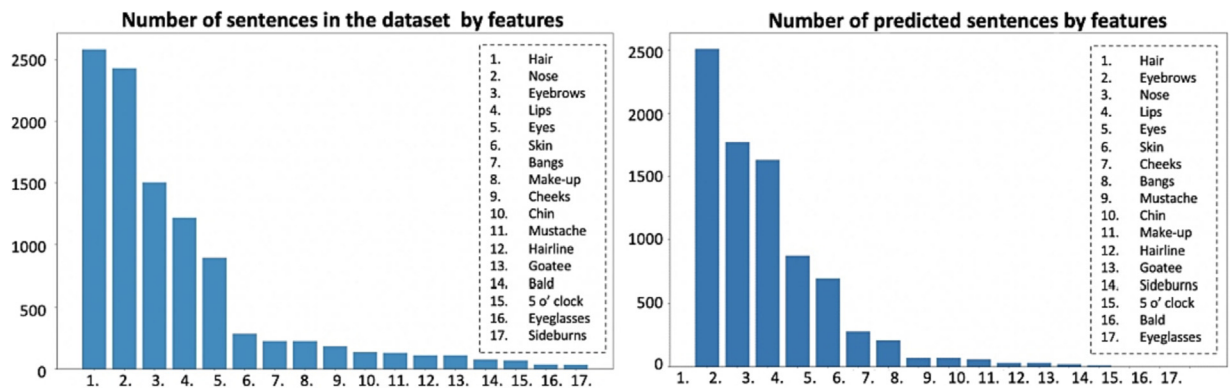

Fig. 7. Distribution of features between the predicted sentences and those in the training dataset:from the most frequent to the least.

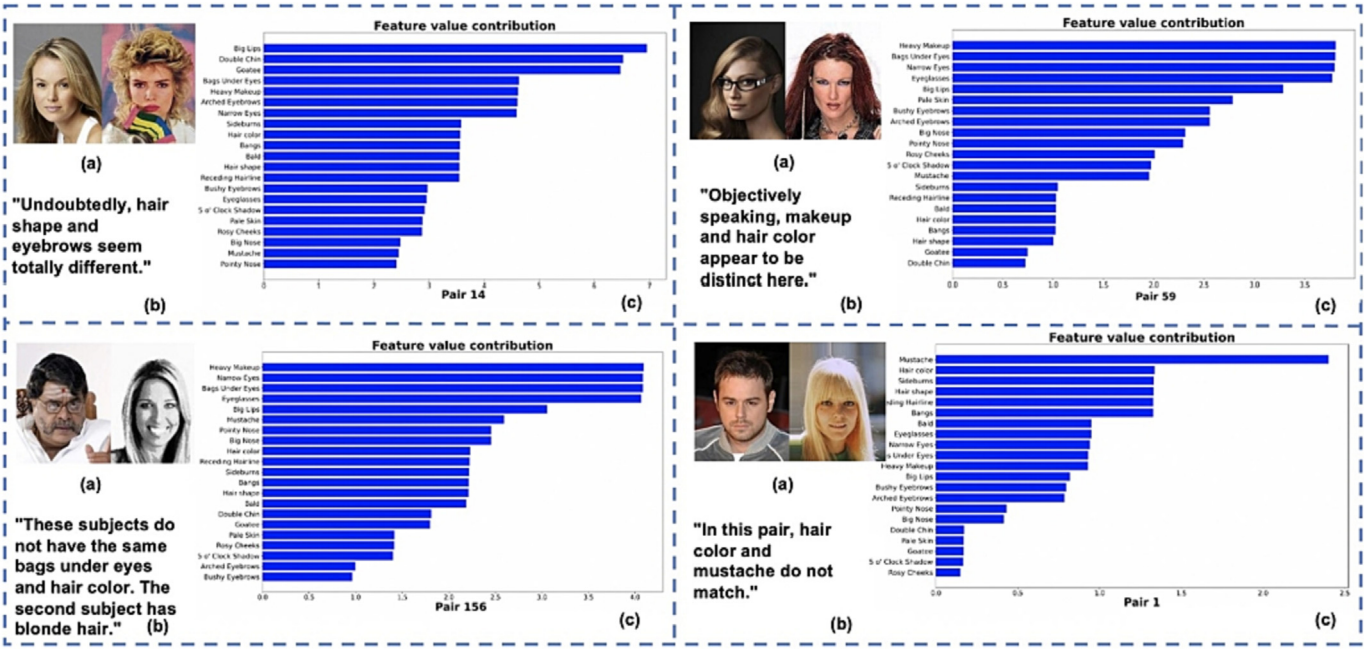

Fig. 8. The figure shows (a) a pair of images present in the test dataset, (b) the caption generated by the text prediction model, (c) an ordered graph that shows the characteristics for which the two subjects differ the most.

6. Conclusion

This paper describes a post-hoc explainability framework to help users comprehend decisions made by a system that assesses the mismatch between two identities using facial attributes. The key objective is to be able to explain why two images seem to be different by generating textual explanations that are as easy to understand as possible while also providing an interpretable mapping between the semantic space of the text and the space of the image. Combining CNN's feature extraction abilities with an LSTM's text generation allows for automatic image captioning. This approach takes a pair of images as input and produces text descriptions that highlight the various facial characteristics. The proposed framework enables the visualization of the distance between the 21 examined facial attributes and objects in a two-dimensional space. Then, using the UMAP strategy, it is feasible to project the facial characteristics into space and visualize them as a point, which facilitates human comprehension by providing a simple and natural explanation. The experiments show the relationship between the text space and the image space, as well as the potential of the two systems, despite the limitations of the dataset.

Authors contribution

Conceptualization, L.C., C.P., and H.P.; methodology, L.C., C.P., and H.P.; software, L.C., C.P., and H.P.; validation, L.C., C.P., and H.P.; formal analysis, L.C., C.P., and H.P.; investigation, L.C., C.P., and H.P.; resources, L.C., C.P., and H.P.; data curation, L.C., C.P., and H.P.; writing—original draft preparation, L.C., C.P., and H.P.; writing—review and editing, L.C., C.P., and H.P.; visualization, L.C., C.P., and H.P.; supervision, L.C., C.P., and H.P.; project administration, L.C., C.P., and H.P.

Data availability

Data will be made available on request.

Declaration of Competing Interest

None.

Acknowledgements

The contributions due to Hugo Proença in this work were funded by FCT/MCTES through national funds and co-funded EU funds under the project UIDB/EEA/50008/2020.

References

[1] A. F. Abate, L. Cimmino, F. Narducci, C. Pero, Biometric face recognition based on landmark dynamics, 2020 IEEE Intl Conf on dependable, autonomic and secure computing, Intl Conf on pervasive intelligence and computing, Intl Conf on cloud and big data computing, Intl Conf on cyber science and technology congress (DASC/PiCom/CBDCom/CyberSciTech), IEEE, 2020, pp. 601–605.

[2] C. Molnar, G. Casalicchio, B. Bischl, Interpretable machine learning–a brief history, state-of-the-art and challenges, Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer 2020, pp. 417–431.

[3] X. Zheng, Y. Guo, H. Huang, Y. Li, R. He, A survey of deep facial attribute analysis, Int. J. Comput. Vis. 128 (8) (2020) 2002–2034.

[4] N. Kumar, A.C. Berg, P.N. Belhumeur, S.K. Nayar, Attribute and simile classifiers for face verification, 2009 IEEE 12th International Conference on Computer Vision, IEEE 2009, pp. 365–372.

[5] T. Berg, P.N. Belhumeur, Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation, 2013 IEEE Conference on Computer Vision and Pattern Recognition 2013, pp. 955–962, https://doi.org/10.1109/CVPR.2013.128.

[6] J. Chung, D. Lee, Y. Seo, C.D. Yoo, Deep attribute networks, Deep Learning and Unsupervised Feature Learning NIPS Workshop, Vol. 3, 2012, p. 1.

[7] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, 2015 IEEE International Conference on Computer Vision (ICCV) 2015, pp. 3730–3738, https://doi.org/10.1109/ICCV.2015.425.

[8] J. Wang, Y. Cheng, R.S. Feris, Walk and learn: facial attribute representation learning from egocentric video and contextual data, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016, pp. 2295–2304, https://doi.org/10.1109/CVPR.2016.252.

[9] Y. Zhong, J. Sullivan, H. Li, Leveraging mid-level deep representations for predicting face attributes in the wild, 2016 IEEE International Conference on Image Processing (ICIP) 2016, pp. 3239–3243, https://doi.org/10.1109/ICIP.2016.7532958.

[10] A. Rozsa, E.M. Rudd, T.E. Boult, Adversarial diversity and hard positive generation, Proceedings of the IEEE conference on computer vision and pattern recognition workshops 2016, pp. 25–32.

[11] A.F. Abate, P. Barra, S. Barra, C. Molinari, M. Nappi, F. Narducci, Clustering facial attributes: narrowing the path from soft to hard biometrics, IEEE Access 8 (2019) 9037–9045.

[12] A. Sethi, M. Singh, R. Singh, M. Vatsa, Residual codean autoencoder for facial attribute analysis, Pattern Recogn. Lett. 119 (2019) 157–165.

[13] N. Zhuang, Y. Yan, S. Chen, H. Wang, C. Shen, Multi-label learning based deep transfer neural network for facial attribute classification, Pattern Recogn. 80 (2018) 225–240.
[14] M. Duan, K. Li, K. Li, Q. Tian, A novel multi-task tensor correlation neural network for facial attribute prediction, ACM Transactions on Intelligent Systems and Technology (TIST), 12 (1), 2020, pp. 1–22.
[15] M. Berrahal, M. Azizi, Augmented binary multi-labeled cnn for practical facial attribute classification, Indones. J. Electr. Eng. Comput. Sci. 23 (2) (2021) 973–979.
[16] Z.C. Lipton, The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery, Queue 16 (3) (2018) 31–57.
[17] C. Molnar, Interpretable Machine Learning, Lulu. com, 2020.
[18] J. Brito, H. Proença, A short survey on machine learning explainability: an application to periocular recognition, Electronics 10 (15) (2021) 1861.
[19] D.V. Carvalho, E.M. Pereira, J.S. Cardoso, Machine learning interpretability: a survey on methods and metrics, Electronics 8 (8) (2019) 832.
[20] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. (2001) 1189–1232.
[21] D.W. Apley, J. Zhu, Visualizing the effects of predictor variables in black box supervised learning models, J. Royal Stat. Soc. Series B Stat. Methodol. 82 (4) (2020) 1059–1086.
[22] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" explaining the predictions of any classifier, Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining 2016, pp. 1135–1144.
[23] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems 2017, p. 30.
[24] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, Proceedings of International Conference on Computer Vision (ICCV), 2015.
[25] A.F. Abate, P. Barra, S. Barra, C. Molinari, M. Nappi, F. Narducci, Clustering facial attributes: narrowing the path from soft to hard biometrics, IEEE Access 8 (2020) 9037–9045, https://doi.org/10.1109/ACCESS.2019.2962010.
[26] E.M. Rudd, M. Günther, T.E. Boult, Moon: A mixed objective optimization network for the recognition of facial attributes, European Conference on Computer Vision, Springer 2016, pp. 19–35.
[27] C.-H. Lee, Z. Liu, L. Wu, P. Luo, Maskgan: Towards diverse and interactive facial image manipulation, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
[28] D.E. King, Dlib-ml: a machine learning toolkit, J. Mach. Learn. Res. 10 (60) (2009) 1755–1758 http://jmlr.org/papers/v10/king09a.html.
[29] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint, 2018 arXiv:1802.03426.
[30] J.P. da Cruz Brito, Deep Adversarial Frameworks for Visually Explainable Periocular Recognition, Ph.D. thesis Universidade da Beira Interior, Portugal, 2021.
[31] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.