Combining ArcFace and Visual Transformer Mechanisms for Biometric Periocular Recognition

João Renato Ribeiro Manesco and Aparecido Nilceu Marana

***Abstract*—In the last decades, advances in Biometrics have resulted in the popularization of biometric identification applications in different scenarios. However, biometric recognition techniques can exhibit sub-par performance in undesirable or restricted scenarios. Therefore, there is still a need to investigate better recognition techniques and more appropriate biometric traits. Studies have shown that attention is an important mechanism present in biological vision systems, including the human vision system, that can improve significantly the correct recognition rates in computer vision systems. Studies have also shown that periocular characteristics suffer less from environmental changes than faces in undesirable scenarios, achieving similar performance using only 25% of all the data in the face. Motivated by these findings, this paper proposes a new method for periocular recognition based on attention mechanisms that incorporates a recent ViT architecture together with the ArcFace loss function. Experimental results obtained on UBIPr and FRGC, two popular datasets, showed that the proposed method obtained lower error rates when compared to other state-of-the-art periocular recognition methods, in addition to being able to provide the visualization of attention weights for a better understanding of the most important periocular regions used by the neural network for biometric recognition.**

***Index Terms*—biometrics, ocular recognition, periocular recognition, attention, visual transformers, arcface**

I. Introduction

Biometric recognition applications have become increasingly more relevant in the last decades, aiming to improve practicality in problems related to people authentication and access management to environments and systems [1].

The use of biometrics has advanced over time, such that recent facial recognition techniques are already able to achieve 99.86% of accuracy on the Labeled Faces In the Wild database [2]. The use of face recognition has become favourable in biometric applications since it is the most natural way for humans to identify themselves, and does not require direct interaction with the authentication system, unlike other types of biometric recognition [1].

Recently, biometric techniques can be seen in a diverse set of applications commonly observed in daily life, like bank transactions, and smartphone access. Even with frequent applications, biometric recognition techniques still can suffer from degradation in performance when applied in non-desirable scenarios, like when face recognition doesn't work properly with big changes in the environment or facial occlusions in the face [3].

João Renato Ribeiro Manesco and Aparecido Nilceu Marana are with UNESP - São Paulo State University, Faculty of Sciences e-mail:{joao.r.manesco, nilceu.marana}@unesp.br.

For that reason, new methods aimed at improve the performance degradation factor in biometric applications are constantly being proposed, even combining features obtained from different domains during training, such as images obtained from visible light and thermal sensors [4].

The struggle of finding the appropriate biometric information results in new biometric recognition techniques being constantly explored and seeking to minimize these types of problems. Among them is periocular biometric recognition, which consists in using information from the region containing the eye and its neighbourhood to identify individuals [5].

Periocular recognition emerges as an alternative to iris recognition, which requires the acquisition of iris images to be done in constrained environments, preferably through infrared sensors, thereby demanding a direct interaction with the sensor [6]. Thus, in order to use the discriminatory characteristics of the iris in unconstrained environments, the periocular region characteristics, in addition to those of the iris, need to be used to compose the biometric feature.

With the arrival of the pandemic caused by the new COVID-19 Coronavirus, direct interaction with authentication systems has become less desirable to avoid contagion risks, and the use of masks has proliferated to minimize the spread of the virus [7]. In environments such as this, biometric recognition methods that involve periocular recognition become even more desirable.

Besides the advantages already discussed, the periocular biometric systems have proven to be less susceptible to variations in the environment than facial images [8]. Figure 1 shows a comparison between different perspectives of the face and their extracted periocular regions, one can perceive that the periocular region of the face suffers less from the impact of perspective variation than the face regions.

Since the periocular region is part of the face, it's easy to extract the biometric information from a wide range of distances from the sensor, unlike iris recognition [9]. Also, datasets popularly used for facial recognition can also be employed in this task, since current face detection techniques can extract the eye position of the face [10].

Among the contributions of this paper, we present a new method for periocular recognition combining both the feature extraction capabilities of Visual Transformers recently introduced in the literature with the metric learning technique ArcFace, commonly used in other biometric applications. We believe our method is capable of achieving state-of-the-art results in cross-dataset periocular authentication while providing a visual interpretation of the most discriminative regions of the eye. The architecture of the proposed technique can be seen

Fig. 1. Comparison between different perspectives of an individual's face and their respective periocular regions.

in Figure 2.

This paper is organized as follows: Section II works show state-of-the-art papers related to periocular biometric recognition. Section III discusses the fundamentals of periocular recognition and sets up the problem we are trying to solve, as well as introduces the periocular region cropping procedure. Section IV introduces the theoretical aspects of the method proposed to solve the task. The methodology used in this paper is described in Section V and the results are discussed in Section VI. Finally, Section VII presents our conclusions and future works.

II. Related Works

An important piece of contribution in the area discusses extensively the possible use of periocular recognition as a biometric tool [9]. The analysis is done by looking at the impact of different aspects of the periocular region on the biometrical recognition task, among which we can cite: the definition of the periocular region, the possible use of the information contained in the iris and sclera, and the effect of variations in the environment, including facial expressions.

A work was proposed to evaluate the performance degradation factors of different regions of interest on the UBIPr dataset, using a fusion of traditional image processing features (HOG, LBP and SIFT) [11].

Another traditional approach consists of finding the most discriminative patches in the periocular region so that patches could match instead of the whole eye [12].

Another work reports results obtained using Probabilistic Deformation Models and m-SIFT. In this case, the patch division of the periocular region allows observing the similarity of each region, providing interpretable results through the most significant comparisons [13].

Recent works show that feature extraction through CNNs pre-trained on large datasets like ImageNet, VGG and VGG-Face produce good results in the task of periocular biometric recognition, so it is possible to use a pre-trained convolutional network as a basis for feature extraction and fine-tuning [14], [15].

Another work studies the impact that changes in facial expressions can cause in the periocular region and finds that there is still a significant loss of efficacy for expressions that were not presented during training [16].

Other techniques try to improve periocular recognition through attention-based neural networks, in which, the object is to emphasize pre-determined relevant semantic regions, in order to find good discriminative features, concluding that the iris and the eyebrow are important in periocular recognition so they should be taken into consideration during further analysis [17], [18].

Alternatively, some methods resort to traditional image processing in order to extract periocular features. One author makes an analysis using LBP, SURF, and SIFT as descriptors, comparing the efficacy of periocular recognition to that of face recognition and concludes that, with only the periocular region, which comprises around 25% of the data found in a face, one is able to achieve similar performance to that of face recognition [19].

Aiming to extract useful iris information, some methods work by fusing the naked-eye periocular information with infrared iris images [20]–[23].

One study proposes that the iris and sclera regions should be ignored during the analysis of periocular recognition [24]. To that end, the authors created a dataset through data augmentation, such that, the importance of the iris region is lessened by having many types of irises found within the same class.

One attempt to create a lightweight model for periocular recognition was proposed using Low-bit Quantization on three popular backbone architectures [25]. The results show that similar performance can be achieved with quantization reducing the model's size by a large margin. Another work employs the usage of CRBM networks for feature learning in addition to supervised metric learning aimed at doing feature reweighting [26].

Knowledge Distillation was also employed in periocular recognition through a template-driven method used to transfer the optimal template extraction knowledge from different ResNet architectures to smaller models, achieving a gain of efficacy even in a cross-device evaluation scenario [27].

A method was proposed using 3D attention mechanisms employed on both the visible spectrum images and near-infrared images, obtaining competitive results in multiple datasets and providing the visualization of the attention feature maps [28].

Finally, a framework is proposed to work in the authentication scenario, providing a visual explanation of why a query was denied in the biometric recognition system [29].

Even though there is a great variety of works in the area of Periocular Recognition, there is a lack of evaluation of these methods in undesirable scenarios. In this work, the cross-dataset evaluation of periocular recognition is explored while covering a recent trend in the literature of providing explainable information during biometric authentication. This is done through the usage of attention maps obtained from a state-of-the-art ViT architecture combined with the ArcFace loss in order to improve the class separability of different subjects.

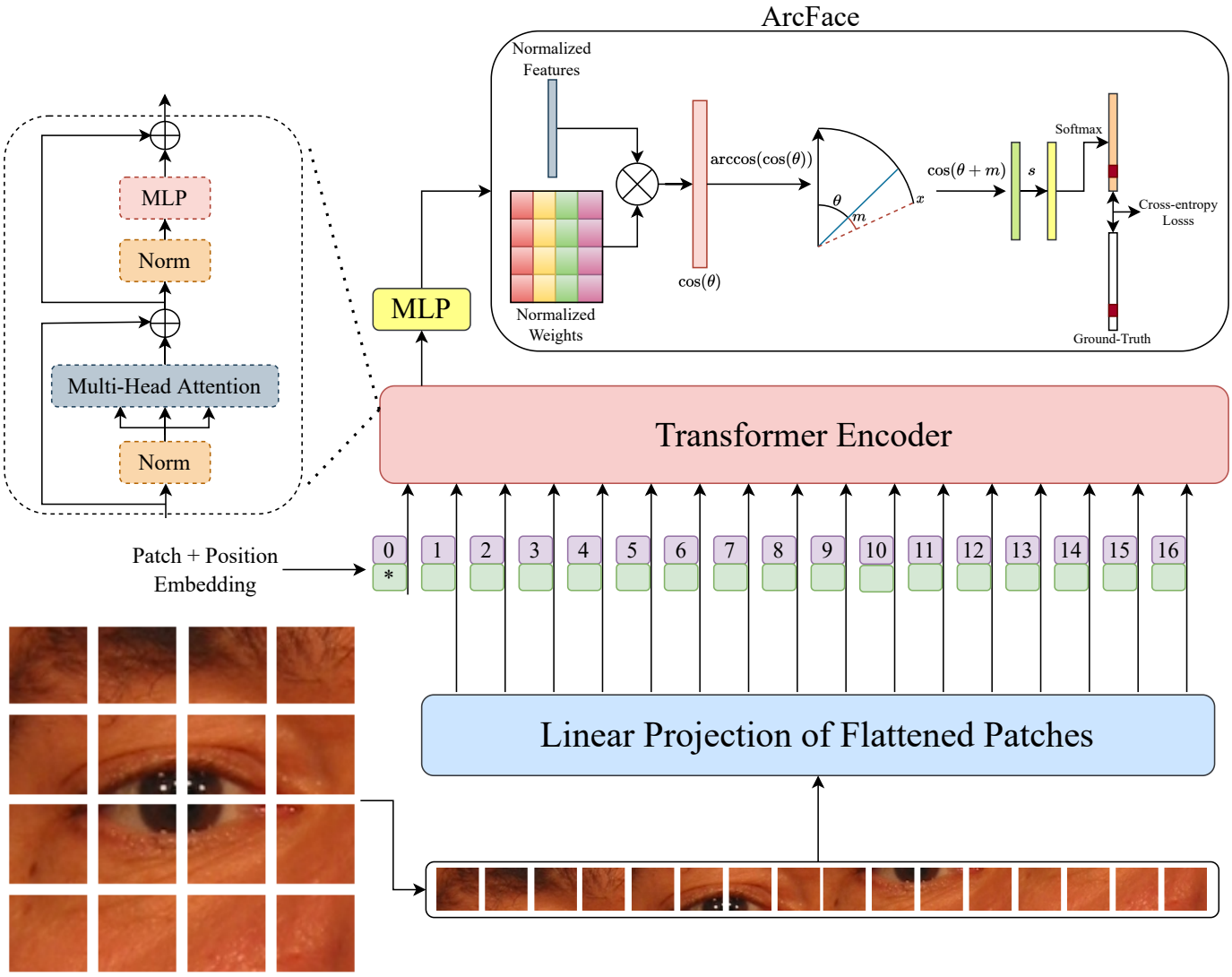


Fig. 2. Pipeline of the proposed method, describing both the transformer step and the ArcFace technique used to position the periocular descriptors near each other in the latent space.

III. Periocular Biometric Recognition

As already established, periocular biometrical recognition consists of using the iris and the area around it as a biometrical feature, used for people identification.

Biometric systems usually operate in two distinct scenarios: authentication and identification. In the authentication task, a similarity measure is established between two samples in order to evaluate if they correspond to the same individual. In identification on the other hand, given an input sample, the objective is to identify which individual better matches that sample [1].

In order to establish the identity of individuals, proper segmentation of the periocular region is required, enabling the extraction of relevant features used to represent the periocular region in the classification process. Therefore, detection and feature extraction steps are imperative to proper biometric recognition.

A. Periocular Region Detection

Periocular region detection consists of defining the area of an image that makes up the ocular region, this is a fundamental process for proper feature extraction. The region detection was discussed before [9], emphasizing the importance of aligning periocular images within the same coordinate system for its proper alignment and scaling, to this end, it relies on information extracted from the iris location.

This subject is further explored by [11], which states that even though the iris can be used to define the scale on images acquired from different camera distances, the alignment should be based on the corner-of-eye information, since the iris can be positioned on other parts of the sclera, causing an offset in the center of the image.

Previously discussed studies work strictly with the segmentation of eye images already extracted from the face region, however, in order to take advantage of a larger number of available images it is interesting to obtain eye regions from facial databases, which are usually available in larger

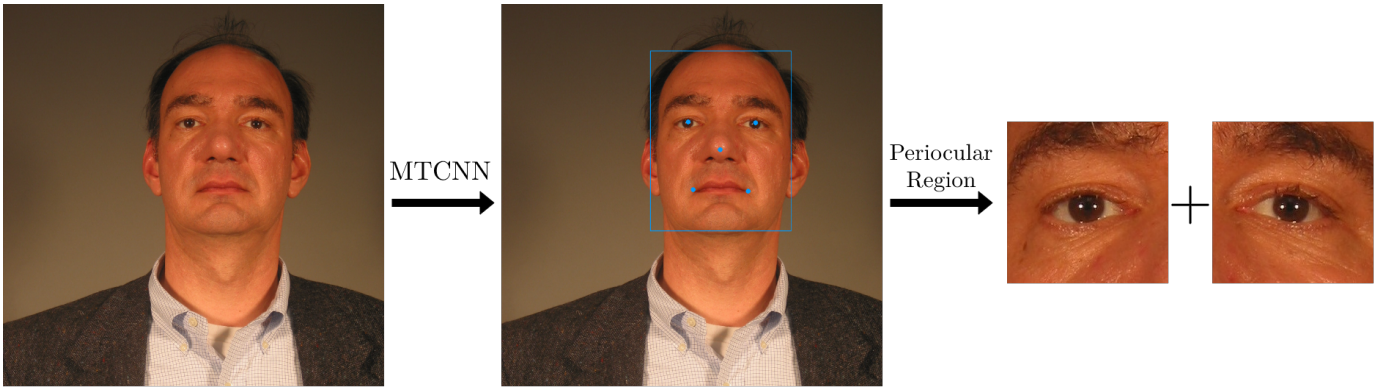


Fig. 3. Periocular region detection through the MTCNN technique. First, the method is applied to obtain the face region and the eye fiducial points, then the ocular region is extracted from the region around the fiducial points.

quantities and cover a larger number of scenarios. Therefore, face detection methods that can detect the fiducial position of the eye, such as the Multitask Cascaded Convolutional Neural Networks (MTCNN) [30], can be used for ocular region extraction.

The MTCNN method is a face detection technique that consists of three cascaded convolutional neural networks used with the objective of detecting possible face regions by successively filtering the regions detected by each network. At the end of the process, the method is also able to detect fiducial points on the face, relative to the nose, mouth, and eyes, obtained with the goal of aiding facial alignment.

It can be used for periocular segmentation by making the fiducial points of the eyes the center of the periocular region, in that way, a simple crop of the region around the eyes is enough to extract the desired information. Figure 3 shows the MTCNN-based periocular detection pipeline. After face detection, the fiducial points are then used to extract each of the detected eyes.

B. Feature Extraction

After the periocular region segmentation, it is necessary to extract relevant features of the periocular region to properly identify the individuals. Paired with the recent popularity of transformer-based techniques in natural language processing, a technique aimed at introducing transformer concepts in computer vision was proposed, showing remarkable performance.

Therefore, in order to evaluate if the performance gain obtained by Visual Transformers (ViTs) [31] in other areas of computer vision could be achieved in periocular recognition, in this work, the ViT was coupled with the ArcFace error function, a popular loss function in biometrics which aims to minimize the inter-class angular distance while the intra-class angular distance is maximized.

On top of providing good performance in other areas of computer vision, there are a few mechanisms proposed to visualize the attention layers in transformer-based architectures, providing a visual interpretation on what are the most relevant features considered by the neural network [32].

IV. Proposed Method

A. Visual Transformers

The visual transformer builds upon the transformer concept initially proposed for natural language processing [31]. The transformer model was employed to solve the machine translation problem following developments in the literature regarding the usage of self-attention [33].

Transformers are sequence-to-sequence models based on an encoder-decoder architecture, in which the input phrases will be tokenized and encoded, together with their respective positional information [34]. The self-attention mechanisms in the transformer architecture, composed of three components: Query, Key and Value, will be responsible to learn the relationship between elements of the sequence [35]. Visual transformers expand on the transformer concept by splitting an image into $n \times n$ patches before introducing them to the encoder.

B. ArcFace

In this work, the Arcface error function is employed during training [36], a popular choice for face recognition. The objective of the feature extraction, in this case, is to jointly minimize the angular distance between elements of the same class while maximizing the inter-class similarity. To this intent, the error function represented by Equation 1 was proposed.

$$L = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))} + \sum_{j=1,j\neq y_i}^{n} e^{s\cos(\theta_j)}} \quad (1)$$

In this case, the samples are represented by a hypersphere with radius s, and the objective of the error function is to minimize the angular distance between samples on this hypersphere, causing the features extracted in the last layer, prior to softmax, to represent a sample in this new space, in which, similar samples are located between an angular distance interval m.

V. Materials and Methods

In order to achieve a better understanding of the methods and evaluate their generalization capabilities, two datasets were used during the experiments: FRGC and UBIPr.

The FRGC dataset [3] was proposed to help explore different challenges in the area of face recognition. The dataset is composed of images from two different settings: constrained and unconstrained. The constrained scenario contains images in well controlled environments, regarding background and lightning. The unconstrained scenario on the other hand contains images obtained in different environments, far from the capture device, with varying degrees of background complexity, lightning and facial expressions. Figure 4 shows examples of images from this dataset.

Fig. 4. Facial images from the FRGC dataset.

For the FRGC dataset the faces were cropped using the MTCNN technique, as described previously in Section III-A, after proper face alignment. The ocular region was empirically defined for each type of scenario following similar patterns as the ones described by [15].

As the capture distance is different in both scenarios, changes in facial expressions or in placement could impact on the results. The empirical analysis was done by defining a window of 122×122 pixels for the first individual of the dataset. The periocular region window of the remaining subjects was adjusted proportionally based on the proportion of the distance between their eyes and the eyes of the first subject.

All the periocular regions were resized to 224×224 pixels as the standard input normally used on neural networks. Examples of periocular images foud in the FRGC dataset can be seen on Figure 5.

(a) Controlado (b) Não Controlado (c) Não Controlado

Fig. 5. Periocular regions of a singular subject found in the different scenarios of the FRGC dataset.

The UBIPr dataset [11] is a periocular recognition dataset composed of 10,252 periocular images of 344 subjects. The images were obtained in different environments ranging from capture distance, face rotation and expression. The dataset is composed of two sessions obtained in distinct days. Figure 6 shows examples of images from this dataset.

A. Experimental Protocol

The pre-processed FRGC dataset information was used to train the proposed architecture, meanwhile, the results were evaluated on the UBIPr dataset in order to observe cross-dataset generalization, providing a comparison in which different sensors and scenarios were used. The method was evaluated through standard biometric authentication metrics, such as Equal Error Rate (EER) and the area under the ROC curve [1].

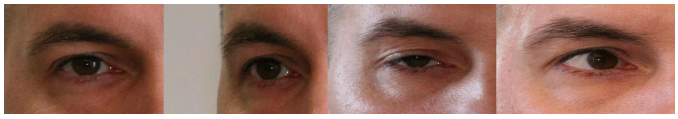


Fig. 6. Left-eye samples of a single subject from the UBIPr dataset considering distinct scenarios.

As there are several approaches for evaluating and comparing periocular techniques, our method was compared to other techniques that perform evaluations on the UBIPr dataset. In our particular case, the method also includes a cross-dataset evaluation as it was trained on a different dataset.

The experiments were conducted using a computer with two Intel Xeon E5620 CPUs, 48GB of RAM, and an NVIDIA TitanXP GPU with 12GB of VRAM, with a batch size of 256, a learning rate of $1e-4$ paired with the Adam optimizer. For the Arcface parameters, the embedding size was 1024 and the standard $m = 0.5$ and $s = 64$ were chosen.

VI. Results

The proposed technique was evaluated in the open-set authentication scenario considering a cross-dataset analysis to evaluate the generalization capabilities of the method. Figures 7 and 8 show the Receiver Operating Characteristic (ROC) curves of the results when the proposed method was trained on the FRGC and evaluated on the UBIPr and when trained on UBIPr and evaluated on the FRGC respectively.

From the experiments, it is possible to observe that the full FRGC training set seems to have more generalization capabilities, being able of obtaining good results when applied to the UBIPr dataset, obtaining 98.79% of the area under the curve (AUC), in contrast to the 77.44% of AUC obtained when the method is trained using the UBIPr dataset and applied to the FRGC dataset.

Table I shows the Equal Error Rate (EER) values obtained by the proposed method (ViT+ArcFace) and by other state-of-the-art (sota) methods of literature when applied to the UBIPr dataset. One can observe that the proposed method showed the best result. When compared with the best sota method, there was a decrease in the EER value from 6.40% to 5.41%.

On top of being competitive, the attention mechanism found in the ViT architecture offers a way to have insights in the areas of analysis of the proposed network when extracting features.

Figure 9 shows the mean attention heatmap on top of the input image, on both the FRGC test dataset (upper row) and the UBIPr dataset (lower row). As one can see, the network tends to focus on the periocular region of the face to make decisions, looking into the eyes when necessary, this is helpful in situations where the ocular region is blurred making it difficult to observe the region of the iris.

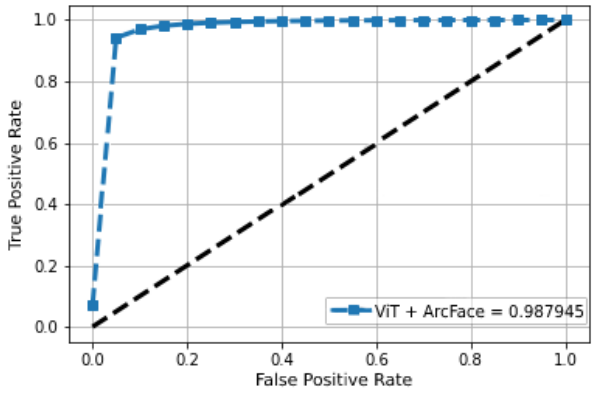


Fig. 7. ROC Curve of the cosine similarity of the UBIPr dataset when the model was trained with FRGC dataset.

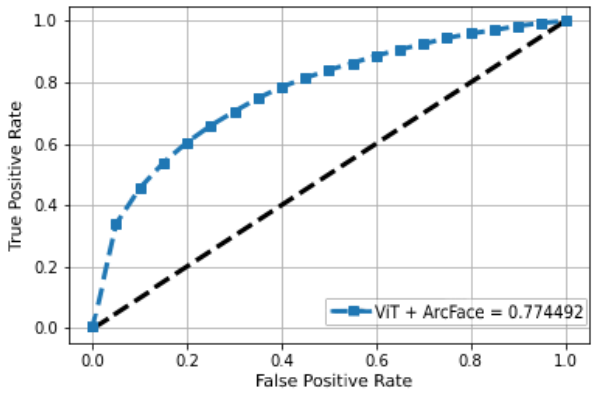


Fig. 8. ROC Curve of the cosine similarity of the FRGC dataset when the model was trained with UBIPr dataset.

TABLE I
EQUAL ERROR RATE VALUES OBTAINED BY VIT+ARCFACE AND OTHER STATE-OF-THE-ART METHODS DURING THE AUTHENTICATION TASK IN THE UBIPR DATASET.

Method	EER
[11]	20.00%
[12]	07.21%
[13]	06.43%
[26]	06.40%
ViT+ArcFace	05.41%

Using the hardware configuration and the parameters described in Section V-A, the training step of our method took 86 seconds, on average, per epoch, to execute (in our experiments we have used 200 epochs). The feature extraction step took 163ms per batch and the matching step took 40ms. We did not present the processing times of the other methods used for comparison in this work because their authors did not provide this information in their papers.

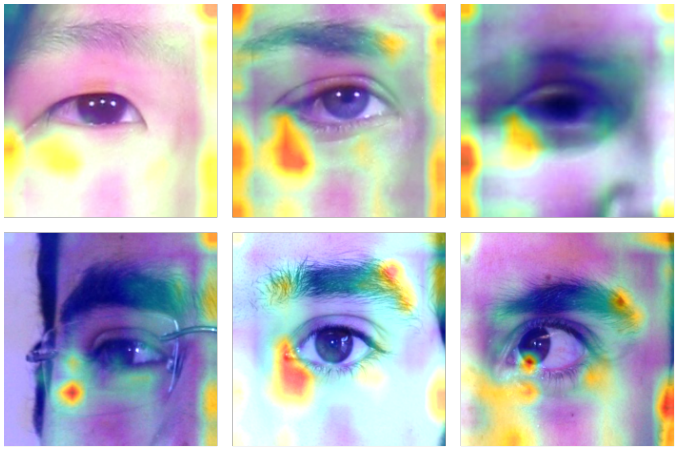


Fig. 9. Mean attention maps obtained on both the evaluated datasets FRGC (first row) and UBIPr (second row).

VII. CONCLUSIONS

The periocular recognition method proposed in this paper (ViT+ArcFace) aims to combine the feature extraction power of a pre-trained ViT with the power of the metric learning technique ArcFace. From the experimental results, it is possible to conclude that using ArcFace in combination with the feature extraction capabilities of the recent ViT models is beneficial for periocular biometric recognition since they reduce the error rate. The proposed method was able to achieve powerful and competitive results obtaining lower error rates when compared to other state-of-the-art periocular recognition methods.

It is also important to notice that ViT models offer a way to visualize the attention weights of each batch of the network, giving insight into the decision process of the proposed method. In our experiments, it was evidenced that the network focused mostly on the periocular regions of the face instead of focusing on the eye region. This can be particularly helpful in situations where the eye regions are occluded or blurred.

Even though our method offers a gain of efficacy in a cross-dataset scenario, it is still possible to perceive that the attention mechanism doesn't consider too much the iris recognition for the extraction of biometric features. As this area contains useful biometric information, future work must be done to overcome this limitation by employing the fusion of biometric characteristics using distinct attention maps for each region in order to keep the interpretability capabilities of the task.

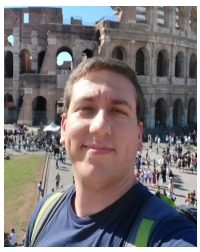
ACKNOWLEDGMENT

The authors thank FAPESP (Process: 2021/02028-6 and 2022/07055-4) and Petrobras/Fundunesp (Process 2662/2017) for the financial support. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES). The TitanXp used for this research was donated by the NVIDIA Corporation.

REFERENCES

[1] A. K. Jain, A. A. Ross, and K. Nandakumar, *Introduction to biometrics*. New York, NY: Springer Science & Business Media, Nov. 2011.

[2] G. Guo and N. Zhang, "A survey on deep learning based face recognition," *Computer Vision and Image Understanding*, vol. 189, p. 102805, 2019.
[3] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, and W. Worek, "Preliminary face recognition grand challenge results," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, (Southampton, UK), pp. 15–24, IEEE, 2006.
[4] N. K. Benamara, E. Zigh, T. B. Stambouli, and M. Keche, "Towards a robust thermal-visible heterogeneous face recognition approach based on a cycle generative adversarial network," *Int. J. Interact. Multimed. Artif. Intell*, vol. 7, no. 4, pp. 132–145, 2022.
[5] P. Kumari and K. Seeja, "Periocular biometrics: A survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 4, pp. 1086–1097, 2022.
[6] L. A. Zanlorensi, R. Laroca, E. Luz, A. S. Britto Jr, L. S. Oliveira, and D. Menotti, "Ocular recognition databases and competitions: A survey," *Artificial Intelligence Review*, vol. 55, no. 1, pp. 129–180, 2022.
[7] J. Howard, A. Huang, Z. Li, Z. Tufekci, V. Zdimal, H.-M. Van Der Westhuizen, A. Von Delft, A. Price, L. Fridman, L.-H. Tang, *et al.*, "An evidence review of face masks against covid-19," *Proceedings of the National Academy of Sciences*, vol. 118, no. 4, p. e2014564118, 2021.
[8] G. Santos and H. Proença, "Periocular biometrics: An emerging technology for unconstrained scenarios," in *2013 IEEE Symposium on Computational Intelligence in Biometrics and Identity Management (CIBIM)*, pp. 14–21, 2013.
[9] U. Park, R. R. Jillela, A. Ross, and A. K. Jain, "Periocular biometrics in the visible spectrum," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 1, pp. 96–106, 2011.
[10] J. Xiang and G. Zhu, "Joint face detection and facial expression recognition with mtcnn," in *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, pp. 424–427, IEEE, 2017.
[11] C. N. Padole and H. Proenca, "Periocular recognition: Analysis of performance degradation factors," in *2012 5th IAPR international conference on biometrics (ICB)*, pp. 439–445, IEEE, 2012.
[12] J. M. Smereka, B. V. Kumar, and A. Rodriguez, "Selecting discriminative regions for periocular verification," in *2016 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, pp. 1–8, IEEE, 2016.
[13] J. M. Smereka and B. V. Kumar, "What is a" good" periocular region for recognition?," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 117–124, IEEE, 2013.
[14] M. V. Vizoni and A. N. Marana, "Ocular recognition using deep features for identity authentication," in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pp. 155–160, 2020.
[15] K. Hernandez-Diaz, F. Alonso-Fernandez, and J. Bigun, "Periocular recognition using cnn features off-the-shelf," in *2018 International conference of the biometrics special interest group (BIOSIG)*, pp. 1–5, IEEE, 2018.
[16] R. C. Dalapicola, R. T. V. Queiroga, C. T. Ferraz, T. T. N. Borges, J. H. Saito, and A. Gonzaga, "Impact of facial expressions on the accuracy of a cnn performing periocular recognition," in *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pp. 401–406, IEEE, 2019.
[17] 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*, vol. 12, no. 5, pp. 1017–1030, 2016.
[18] Z. Zhao and A. Kumar, "Improving periocular recognition by explicit attention to critical regions in deep neural network," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 12, pp. 2937–2952, 2018.
[19] K. K. Kamarajugadda and P. Movva, "Periocular region based biometric identification using sift and surf key point descriptors," in *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 0968–0972, IEEE, 2019.
[20] S. Umer, A. Sardar, B. C. Dhara, R. K. Rout, and H. M. Pandey, "Person identification using fusion of iris and periocular deep features," *Neural Networks*, vol. 122, pp. 407–419, 2020.
[21] G. Santos, E. Grancho, M. V. Bernardo, and P. T. Fiadeiro, "Fusing iris and periocular information for cross-sensor recognition," *Pattern Recognition Letters*, vol. 57, pp. 52–59, 2015.
[22] F. Boutros, N. Damer, K. Raja, R. Ramachandra, F. Kirchbuchner, and A. Kuijper, "Iris and periocular biometrics for head mounted displays: Segmentation, recognition, and synthetic data generation," *Image and Vision Computing*, vol. 104, p. 104007, 2020.
[23] D. K. Jain, X. Lan, and R. Manikandan, "Fusion of iris and sclera using phase intensive rubbersheet mutual exclusion for periocular recognition," *Image and Vision Computing*, vol. 103, p. 104024, 2020.
[24] 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*, vol. 13, no. 4, pp. 888–896, 2017.
[25] J. N. Kolf, F. Boutros, F. Kirchbuchner, and N. Damer, "Lightweight periocular recognition through low-bit quantization," in *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–12, IEEE, 2022.
[26] L. Nie, A. Kumar, and S. Zhan, "Periocular recognition using unsupervised convolutional rbm feature learning," in *2014 22nd International Conference on Pattern Recognition*, pp. 399–404, IEEE, 2014.
[27] F. Boutros, N. Damer, K. Raja, F. Kirchbuchner, and A. Kuijper, "Template-driven knowledge distillation for compact and accurate periocular biometrics deep-learning models," *Sensors*, vol. 22, no. 5, p. 1921, 2022.
[28] S. S. Behera and N. B. Puhan, "High boost 3-d attention network for cross-spectral periocular recognition," *IEEE Sensors Letters*, vol. 6, no. 9, pp. 1–4, 2022.
[29] J. Brito and H. Proença, "A deep adversarial framework for visually explainable periocular recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1453–1461, 2021.
[30] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, pp. 1499–1503, Oct 2016.
[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[32] A. Steiner, A. Kolesnikov, , X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, "How to train your vit? data, augmentation, and regularization in vision transformers," *arXiv preprint arXiv:2106.10270*, 2021.
[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), p. 6000–6010, Curran Associates Inc., 2017.
[34] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, vol. 3, pp. 111–132, 2022.
[35] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.
[36] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.



João Renato Ribeiro Manesco is an M.Sc. student at the São Paulo State University (UNESP), Brazil. He received his Bachelor's degree in 2020 at the same University. In 2022, he did an international research internship at the Media Integration and Communication Center at the University of Florence, Italy. His research interests include Biometrics, 3D Computer Vision, Human Pose Estimation and Domain Adaptation.

Dr. Aparecido Nilceu Marana graduated in Mathematics from São Paulo State University – UNESP - Brazil (1985). He holds a Master's degree in Computer Science from the State University of Campinas – UNICAMP (1990) and a PhD in Electrical Engineering also from the State University of Campinas – UNICAMP (1997) – Brazil. In 1996 and 2005 he was a visiting scholar at King's College London, UK, and Michigan State University, USA, respectively. Since 2005, he works in the field of Biometrics. His research interests also include Computer Vision, Image Processing, Pattern Recognition and Machine Learning. He is currently an Associate Professor at the São Paulo State University (UNESP) and coordinates, together with Dr. João Paulo Papa, the Recogna Research Laboratory. Since 1989 he has been a member of the Brazilian Computer Society.