2024 IEEE 12th International Conference on Serious Games and Applications for Health (SeGAH) | 979-8-3503-8438-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/SeGAH61285.2024.10639602

Eye Importance in Facial Expression Recognition

Ana Sofia Rodrigues
Research Center for Digitalization and Intelligent Robotics, Instituto Politécnico de Bragança
Email: ana-rodrigues@ipb.pt

Júlio Castro Lopes
CeDRI, Instituto Politécnico de Bragança, Portugal
SusTEC, Instituto Politécnico de Bragança, Portugal
Email: juliolopes@ipb.pt

Rui Pedro Lopes
CeDRI, Instituto Politécnico de Bragança, Portugal
SusTEC, Instituto Politécnico de Bragança, Portugal
Email: rlopes@ipb.pt

***Abstract*—The human face is a powerful tool for nonverbal communication, capable of conveying a wide range of emotions. Previous research has shown that facial expressions contribute significantly to interpersonal communication, indicating that 55% of information is conveyed through facial expression alone. Despite advancements, Facial Expression Recognition (FER) technology faces challenges, particularly in scenarios involving occlusions. Instances during the COVID-19 pandemic and in Virtual Reality (VR) environments highlight these challenges, where mask usage and head-mounted displays obstruct facial features critical for accurate recognition. This paper aims to investigate the importance of the eyes in facial expression recognition by using four models: ResNet-18, VGG-19, EfficientNet-B1, and an Ensemble model. Utilising the FERPlus dataset, scenarios with and without occlusion were examined. In scenarios without occlusion, ResNet-18 emerged as the top-performing model, achieving 86.1% accuracy. However, when occluded by goggles, the Ensemble model demonstrated superior performance with 82.8% accuracy. Furthermore, in the presence of mask occlusion, EfficientNet-B1 exhibited the most robust performance, achieving an accuracy of 71.2%. Despite challenges, the results of this paper reaffirm the enduring importance of the eyes in facial expression recognition, emphasizing their pivotal role in conveying emotions even amidst technological obstacles.**

***Index Terms*—Facial Expression Recognition, Partial Occlusion, Virtual Reality, Eyes, COVID-19, Serious Games**

I. Introduction

The human face stands out as the epitome of effective nonverbal communication, attributed to its complexity and significance. Serving as a dynamic channel of nonverbal expression, it possesses the ability to convey both involuntary reactions and deliberate gestures, rendering it a highly significant feature in interpersonal communication [1]. Mehrabian [2] discovered that 75% of the information is communicated between individuals through writing, 38% through conversation, and 55% through facial expressions.

Facial expression recognition (FER) has received significant attention in diverse fields, such as computer vision, psychology, and human-computer interaction [3]. Ekman and Friesen [4] identified a set of facial expressions of emotion that transcend cultural boundaries: surprise, fear, disgust, anger, happiness, and sadness. Subsequently, the same researchers created a facial atlas detailing the involvement of each muscle in the six fundamental emotional states, providing the foundation for an encoding system [5].

In the past few years, the application of FER technology has expanded its reach to the serious game industry, achieving a significant breakthrough in exploiting the capabilities of FER for immersive gaming experiences. Through this technology, interactions between in-game characters or scenarios can be more immersive and engaging, as emotions can be dynamically responded to, leading to a more personalized and captivating gaming experience [6]. Moreover, it can contribute to therapeutic interventions, including Virtual Reality (VR) exposure therapy for anxiety disorders [7] and educational [8].

Despite the advancements, this technology still faces challenges in accuracy and reliability, especially when dealing with occlusions. An illustrative instance lies in its application during the COVID-19 pandemic, where widespread mask usage obscured everything below the nose, including the nose itself, thereby complicating the recognition of emotions [9]. Another example arises when this technology is applied to VR environments, particularly when a person is using a head-mounted display, which occludes the area around the eyes, a crucial region for accurate facial expression recognition [10].

The primary goal of this paper is to investigate the role of the eyes in FER by performing standard FER and comparing it with two distinct scenarios in which a portion of the face is occluded (Fig. 1). The first occlusion scenario (Fig. 1a) involves FER with the upper half of the face occluded by VR goggles. The second scenario (Fig. 1b) involves performing FER while wearing facial masks, that cover the lower half of the face. As a result of these simulations, it is expected that a deeper understanding of the role of the eyes in recognizing facial expressions will be gained. The design and implementation of more sophisticated gaming systems can be influenced by understanding the pivotal role of the eyes in conveying emotions. For instance, incorporating mechanisms that react to subtle changes in facial expressions can enable games to dynamically respond to emotional states of the players, leading to more personalized and immersive experiences [6]. In addition, these insights can lead to the development of new gameplay mechanics that rely on ocular interactions, which can provide players with innovative ways to interact with virtual environments.

(a) First Scenario (b) Second Scenario

Fig. 1: Occlusion Scenarios

II. Related Work

Up to this moment, the authors of this paper have not found any paper that focuses on studying the real importance of the eyes in FER, which is the focus of the present paper. This section aims to explore and describe recent advances in the classification of facial expressions with partial occlusion of the face.

A system that recognizes facial expressions with face masks was proposed by Yang et. al [11]. The authors proposed a two-step strategy: firstly, they re-trained a face-mask-aware face parsing model using existing datasets with face mask annotations, and secondly, they employed a vision Transformer capable of handling both occluded and non-occluded facial regions. The EHANet [12] model was able to process CelebaMask-HQ [13] and ibugmask [14] datasets in the initial step, leading to the creation of M-CelebAMask-HQ and M-ibugmask datasets. In the subsequent phase, facial images were identified by MTCNN [15] and 2D tensors for various datasets, including, M-KDDI-FER [16], M-CK+ [17] (automatically processed as facial image wearing face) and M-FER2013 [18]. Following this, the researchers utilized the ResNet50 backbone, which had been pre-trained on ImageNet, to extract features from the 2D tensor for each branch. The accuracy on M-LFW-FER, M-KDDI-FER, M-FER2013, and M-CK+ was 90.31%, 91.83%,66.53% and 61.08% respectively.

Magherini et al. [19] also developed a system that can recognize emotions when half of the face is occluded by a face mask. In their research, the AffectNet dataset [20] was used, which has some images with occlusion elements, such as sunglasses and hats. Since this dataset lacked face images with masks, they applied the MaskTheFace algorithm [21], which identifies the tilt of the face and places a mask. The authors developed a framework consisting of two main blocks: the first block validated images via ResNet-50 [22], filtering out occluded elements such as sunglasses, hats, and shadows that obscured the forehead and eye regions, impeding emotion recognition in the presence of face masks, while the second block conducted emotion recognition, employing the Inception network [23]. The framework achieved an accuracy of 96.71%.

Castellano et al. [24] proposed a method to recognize emotions from masked faces by utilizing a Convolutional Neural Network (CNN) that is based on MobileNetV2 [25], trained on ImageNet [26] for mask detection. When a mask was detected, the system focused on the eye area for emotion analysis; otherwise, it analyzed the entire face. Two ResNet-50 [22] models were trained: one for analyzing the full face when no mask was detected and another for focusing on the eye area when a mask was detected. Additionally, the authors tested both models with a Bottleneck Attention Module (BAM) [27], named ResNetBAM. They curated two datasets from FER2013 [18]: FER2013_filtered containing images with detectable facial landmarks and FER2013_cropped for developing the FER model for masked faces. Results showed 63.52% accuracy for ResNet-50 and 62.62% for ResNet50BAM using FER2013_cropped and 73.21% and 74.32% using FER2013_filtered, respectively.

Abate et al. [28] examined the effects on FER systems, of faces that have been obscured by masks, as well as those with obstructed eyes. The authors explored four scenarios: in the first, unmasked faces were used for training and masked faces for testing; the second scenario utilized masked faces for both training and testing; in Scenario 3, a model trained on unmasked faces was tested with eye-occluded images; and Scenario 4 involved training and testing with eye-occluded images. They employed three algorithms (Residual masking network [29], FER CNNs [30] and Amend-Representation Module [31])), trained and tested on datasets including FER2013 [18] and RAF-DB [32]. Additionally, experiments involving masked (with fake masks generated as described in [33]) and eye-occluded (using a simple black occlusion bar placed over the periocular region) conditions were conducted on these two datasets. The highest performance, in terms of accuracy, in all scenarios was attained using the Amend-Representation Module on the RAF-DB dataset: 45.45% in the first scenario, 82.30% in the second, 74.25% in the third and 84.32% in the last scenario.

Ruan et al. [34] developed a strategy that uses a path selection multi-network model to recognize facial expressions in three different facial occlusion scenarios: upper face occlusion, lower face occlusion and eye occlusion. The authors combined multiple expression recognition databases including FER2013 [18], JAFFE [35], KDEF [36], and RAF-DB [32]. The upper and lower halves of images were directly blackened to simulate facial masking. To simulate eye occlusion, they used the Haarcascade_eye.xml classifier to locate both eyes and shade their corresponding areas. New databases were created through this process, including ConcatDB, which contains images with upper face occlusion, BConcatDB, which contains images with lower face occlusion and EyeCDB that consists exclusively of images with eye occlusion. By segmenting the labels within a single database, three new sub-databases were created to train three Subnets (networks developed by the authors) separately. In order to make path selection easier, an integration method was used, that involved multi-networks, where groups of labels were combined in a single database to train an initial network that was labeled beginner. Subsequently, the prediction made by BeginNet (network constructed by the authors) dictated the Subnet responsible for making the final prediction. The accuracy rate for the proposed methodology was 59.8%, 60.6% in BConcatDB, and 63.71% in eyeCBD, respectively.

III. Methodology

To determine the impact of the eyes on FER, the procedure involves assessing three CNN networks (VGG-19, ResNet-18, and EfficientNet-B1) and then assembling them, creating a "stronger" network.

To achieve this, an occlusion algorithm was integrated, which replicates the effect of VR goggles. A mask simulation algorithm was also utilized to mimic surgical masks.

A. Dataset

The FER+ dataset (Figure 2) can be described as an expanded version of the FER2013 dataset [18]. Images were categorised into 8 emotion classes (0 = Neutral, 1 = Happiness, 2 = Surprise, 3 = Sadness, 4 = Anger, 5 = Disgust, 6 = Fear, and 7 = Contempt). The purpose of this dataset was to address problems in FER2013, such as correcting misclassified images and eliminating those with missing facial features. In addition, each image had tags from 10 crowd-sourced annotators, which means that each row was aiming to cover a total of 10 values (the 8 emotions already mentioned, "unknown" and "NF" - not a face). These values are equivalent to the probabilities assigned to different emotions by the annotators. The reclassification process began by ignoring columns labeled as 'unknown' and 'NF' (not a face) since they were not related to our classification objectives. Afterward, the highest probability value for every row was established, assigning numerical values between 0 and 7 to symbolize the 8 emotions. Due to the frequent occurrence of duplicated highest probability values within the same row, manually inspecting all images to determine the most probable emotion would have been both impractical and time-consuming. Therefore, a random selection process was used in these cases. After this process, a new CSV file was created, that has two columns, 'Image name' and 'label', containing the revised labels.

Fig. 2: FER+ dataset samples

To simulate the presence of VR goggles in the FER+ dataset, preprocessing steps were undertaken. This involved utilizing an occlusion algorithm (https://github.com/SofiaRodrigues41737/Occlusion_Algorithm), developed in previous work **pereira_classification_2022**. The algorithm consists of two stages: facial detection and landmark localization using an MTCNN (Multi-task Cascaded Convolutional Networks), followed by determining the position of the goggles. Initially, the MTCNN network was employed to detect the facial structure and identify key facial landmarks, including the eyes, mouth corners (left and right), and nose [15]. Subsequently, based on this information, the position of the VR goggles was computed and added to the original images (Fig. 3).

Fig. 3: Sample of the FER+ dataset with VR occlusion.

To simulate mask occlusion, a masking algorithm was additionally implemented as detailed in [37]. The algorithm is composed of MaskTheFace, a computer vision script, that is designed to conceal faces in images. It employs a dlib-based face landmark detector to identify facial tilt and key facial features required for mask application. After selecting the appropriate mask template based on the face tilt, it precisely tailors the template to conform to the facial contours while also considering factors such as face angle and lighting conditions when applying masks to all detected faces, as shown in Fig. 4.

Fig. 4: Sample of the FER+ dataset with mask occlusion

After the procedure, it was discovered that there were discrepancies in the number of samples for the eight classes in the training set (Fig. 5).

To mitigate the imbalance, a weighted loss function was employed, incorporating a weight parameter within the CrossEntropyLoss function as specified in Eq. 1.

$$loss(x, y) = -weight[y] * \frac{log(exp(x[y]))}{sum(exp(x))} \quad (1)$$

, where x represents the model output, y denotes the target class, exp signifies the exponential function, and $sum(exp(x))$ indicates the sum of exponentials across all classes.

This modified function adjusts the standard loss function by introducing weights, thereby enhancing the sensitivity of the models to minority classes, consequently increasing the penalty for misclassifying such classes. The underlying principle is to assign higher weights to minority classes and lower weights to majority classes.

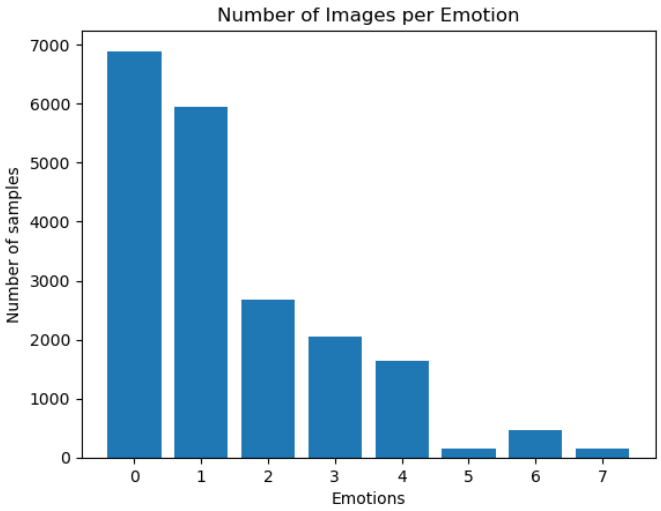

Fig. 5: Original dataset with 8 classes

B. CNNs

In this study, four CNNs were utilized: VGG-19, ResNet-18, EfficientNet-B1, and an ensemble combining these three models.

The VGG (Visual Geometry Group) network, devised by Simonyan and Zisserman at the University of Oxford in 2014, is a well-known convolutional neural network (CNN) model [38]. Trained on the ImageNet ILSVRC dataset with 1.3 million images and 1000 classes, VGG-19, a variant with 19 interconnected layers, has consistently outperformed other models. Its architecture, featuring both convolutional and fully-connected layers, coupled with Maxpooling for downsampling and SoftMax activation for classification, facilitates robust feature extraction.

ResNet-18 (Residual Network 18) [39] is another pretrained model on the ImageNet dataset, consisting of 18 layers including 17 convolutional layers, a fully connected layer, and a softmax layer for classification. The convolutional layers utilize 3x3 filters, and the input size is 224x224x3. Downsampling is achieved through convolutional layers with a stride of 2. The network also incorporates average pooling, followed by a fully connected layer with softmax activation. Residual connections are introduced between layers to facilitate learning.

EfficientNet-B1 is built upon a basic component known as the MBConv module. This module alters feature channels using a 1x1 convolution followed by a depth-wise convolution. Subsequently, it introduces a channel attention mechanism inspired by SENet [40]. Finally, the feature map channels are reduced using a 1x1 convolution.

In the end, an ensemble of the three models was also tested. Given its higher processing and memory requirements, the ensemble was executed using the max voting technique. This method involves collecting the predictions from each model (VGG-19, ResNet-18, and EfficientNet-B1) and selecting the emotion that receives the most votes across all models.

IV. Results

VGG-19 and ResNet-18 underwent training for 100 epochs, while EfficientNet-B1 was trained for 200 epochs, using optimal configurations determined through hyperparameter tuning. All networks were trained using a batch size of 64 samples on a system equipped with a 64GB RAM AMD Ryzen Threadripper 3970X 32-Core Processor and an NVIDIA GeForce RTX 3090.

The FER+ dataset was utilized to train, test, and validate the models, with an 80% split for training, 10% for testing, and 10% for validation. Samples underwent modifications to introduce occlusion over the eyes based on the occlusion algorithm detailed in [10] and to include occlusion of the lower part of the face using the MaskTheFace algorithm [37].

Following preprocessing, experiments were conducted with each mentioned network. Convolutional layers performed operations on input images. The output of the network was compared to ground truth using the loss function specified by Eq. 1. Through backpropagation, gradients of the loss with respect to parameters were calculated to update weights and biases, minimizing the loss. Upon completion of training, images were classified as neutral, happiness, surprise, sadness, anger, disgust, fear or contempt.

Regarding overall accuracy, in the no occlusion scenario, ResNet-18 demonstrated the highest accuracy of 86.1%, followed by the Ensemble with 85.3%, VGG-19 with 83.0%, and EfficientNet-B1 with 82.3% (Table Ia). When simulating the use of VR goggles, the Ensemble model performed best, achieving an accuracy of 82.8%, followed by EfficientNet-B1 with 79.7%, VGG-19 with 79.5%, and ResNet-18 with 77.3% (Table Ib). With mask occlusion, EfficientNet-B1 emerged as the top classifier with an accuracy of 71.4%, followed by VGG-19 with 70.4%, ResNet-18 with 69.5%, and the Ensemble with 68.8% (Table Ic).

In a more detailed analysis, at no occlusion scenario, ResNet-18 exhibited the lowest recognition rates for contempt at 28.0% and fear at 36.6%, while achieving the highest recognition rates for happiness at 92.3% and neutral at 86.1%. In VGG-19, the top two recognized emotions were neutral (86.1%) and happiness (92.3%), whereas the worst performances matched those of ResNet-18, with contempt at 24.0% and fear at 40.0%. Similarly, in EfficientNet-B1, the leading emotions were neutral (86.5%) and happiness (92.9%), while the least recognized emotions mirrored those of the other networks, with contempt at 24.0% and fear at 38.3%. In the case of the Ensemble model, the top two recognized emotions were happiness (94.2%) and surprise (91.4%), while the worst performances were attributed to contempt at 19.0% and disgust at 57.1%.

When upper occlusion was introduced, the recognition rates for emotions with the lowest accuracies in ResNet-18 remained consistent with the previous scenario, staying at 28.0% for contempt and 30.0% for fear. Similarly, in VGG-19 and EfficientNet-B1, this pattern persisted, with contempt at 20.0% and 24.0% respectively, and fear at 38.3% and 35.0% corre-

spondingly. In the Ensemble model, fear and disgust emerged as the worst-performing emotions, with recognition rates of 19.0% and 53.5% respectively. Conversely, the emotions with the highest recognition rates remained consistent with the previous scenario, with ResNet-18 achieving rates of 89.9% and 82.3% in that order. Likewise, in VGG-19, EfficientNet-B1, and the Ensemble model, the top two emotions remained unchanged from ResNet-18, with happiness reaching rates of 91.1%, 92.3%, and 91.8% respectively, while neutral attained rates of 83.6%, 84.7%, and 89.0% correspondingly.

Turning to the scenario involving mask occlusion, in ResNet-18, the lowest recognition rates for emotions persisted at 20.0% for contempt and 30.0% for fear. In VGG-19, EfficientNet-B1, and the Ensemble, they also maintain the trends observed in the previous scenario, with 20.0% for contempt in the first two, and 19.0% in the latter, 40.0% and 33.3% for fear in the first two, and 50.0% for disgust in the Ensemble. However, the highest recognition rates exhibited a shift in ResNet-18; while happiness maintained its position as one of the top two emotions with 75.2%, surprise emerged as the new leading emotion at 78.7%. This shift is also evident in VGG-19 and the Ensemble, with rates of 80.1% and 75.3% for happiness, and 81.2% and 77.9% for surprise, respectively. In EfficientNet-B1, the highest scores are for neutral and happiness, at 76.0% and 81.5% respectively.

To demonstrate the classification capabilities of the models in predicting the eight classes of facial expressions, confusion matrices were employed (Fig. 6, Fig. 7, Fig. 8). These matrices enabled a comparison of results between datasets with occlusion (mask and VR goggles) and without occlusion. In every scenario, it can be noted that fear is commonly mistaken for surprise, occurring in nearly half of the instances. Likewise, contempt is often mistakenly classified as sadness and neutrality, with the majority of instances being classified as neutral, surpassing half of the samples designated as contempt. This occurrence might be attributed to the shared characteristics between different emotions. For instance, between surprise and fear, such as widened eyes, raised eyebrows, and an open mouth. Conversely, in the case of contempt, sadness, and neutrality, there are similarities observed in terms of facial muscle relaxation. Happiness showed notably accurate results, suggesting that the model was successful in discriminating the characteristics associated with this expression.

The findings also indicate that, overall, identifying facial expressions is more challenging when individuals wear masks compared to when they use VR goggles.

Regarding the importance of the eyes in the recognition of facial expressions, despite the observed results, the eyes remain a crucial aspect of facial expression recognition. While masks may partially obscure the mouth region, the eyes often convey a wealth of information about emotions, including happiness, surprise, sadness, and fear. Therefore, even though masks may hinder the visibility of the mouth area, the eyes still play a significant role in conveying and recognizing emotions. Overall, the eyes are integral to facial expression recognition because they convey a wealth of emotional information, facilitate non-verbal communication, and play a significant role in social interaction and understanding. Their importance extends beyond mere visual aesthetics, contributing to our fundamental understanding of human emotions and behavior. Moreover, in serious games, accurate facial expression recognition, particularly focusing on the eyes, enriches the gameplay by enabling dynamic and personalized experiences tailored to emotional states of the players, fostering deeper immersion and engagement.

TABLE I: Accuracy per class with weighted loss function

(a) No occlusion

Class	ResNet-18	VGG-19	EfficientNet-B1	Ensemble
Neutral	0,861	0,861	0,865	0,881
Happiness	0,923	0,923	0,929	0,942
Surprise	0,856	0,865	0,865	0,914
Sadness	0,616	0,662	0,636	0,653
Anger	0,705	0,768	0,685	0,728
Disgust	0,476	0,428	0,428	0,571
Fear	0,366	0,400	0,383	0,576
Contempt	0,280	0,240	0,240	0,190
Accuracy	0,861	0,830	0,823	0,853

(b) Goggles occlusion

Class	ResNet-18	VGG-19	EfficientNet-B1	Ensemble
Neutral	0,823	0,836	0,847	0,890
Happiness	0,899	0,911	0,923	0,918
Surprise	0,818	0,821	0,825	0,856
Sadness	0,498	0,586	0,509	0,555
Anger	0,632	0,666	0,700	0,694
Disgust	0,476	0,523	0,428	0,535
Fear	0,300	0,383	0,350	0,538
Contempt	0,280	0,200	0,240	0,190
Accuracy	0,773	0,795	0,797	0,828

(c) Mask occlusion

Class	ResNet-18	VGG-19	EfficientNet-B1	Ensemble
Neutral	0,725	0,713	0,760	0,694
Happiness	0,752	0,801	0,740	0,753
Surprise	0,787	0,812	0,815	0,779
Sadness	0,536	0,521	0,555	0,521
Anger	0,632	0,570	0,652	0,579
Disgust	0,380	0,380	0,476	0,500
Fear	0,300	0,400	0,333	0,576
Contempt	0,200	0,200	0,200	0,190
Accuracy	0,695	0,704	0,712	0,688

V. Conclusions

This study aims to investigate the significance of the eyes in facial expression recognition. To achieve this, four distinct models were used: ResNet-18, VGG-19, EfficientNet-B1, and an Ensemble model combining these three models. In scenarios without occlusion, utilizing the FERPlus dataset, ResNet-18 emerged as the top-performing model, achieving a notable accuracy of 86.1%. When faced with occlusion caused by goggles, the Ensemble model demonstrated superior performance, reaching an accuracy of 82.8%, as anticipated.

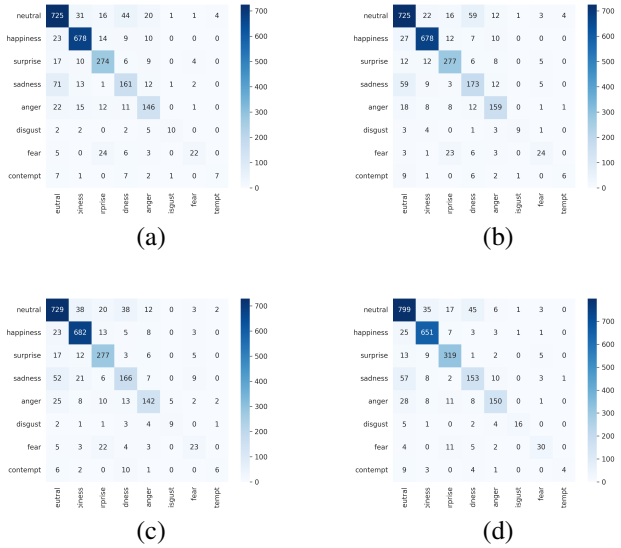

Fig. 6: Confusion Matrices obtained using FERPlus dataset. **(a)** ResNet-18; **(b)** VGG-19; **(c)** EfficientNet-B1; **(d)** Ensemble;

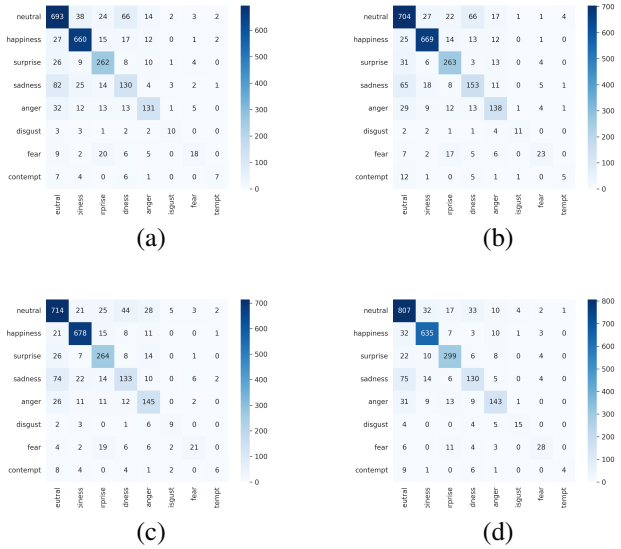

Fig. 7: Confusion Matrices obtained using FERPlus dataset with goggles. **(a)** ResNet-18; **(b)** VGG-19; **(c)** EfficientNet-B1; **(d)** Ensemble;

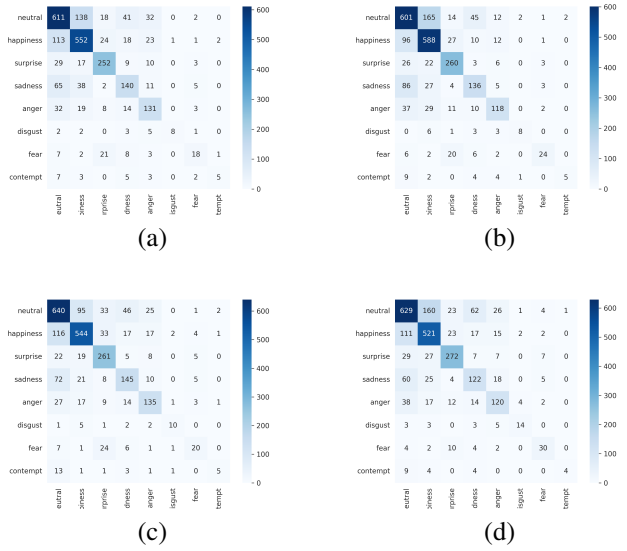

Fig. 8: Confusion Matrices obtained using FERPlus dataset with mask. **(a)** ResNet-18; **(b)** VGG-19; **(c)** EfficientNet-B1; **(d)** Ensemble;

In the case of mask occlusion, EfficientNet-B1 showcased the most robust performance, achieving an accuracy of 71.2%. It is noteworthy that contempt and fear were particularly prone to misclassification, potentially attributed to the significant disparity in sample sizes between these emotions and others within the dataset. It is evident from the results that despite lower accuracies in identifying facial expressions with mask occlusions, the eyes remain pivotal for facial expression recognition. As one of the most vital channels of non-verbal communication, they play a crucial role in conveying emotions even when other facial features are obscured. This underscores the enduring significance of the eyes in understanding and interpreting human expressions, reaffirming their indispensable role in the realm of facial expression recognition.

VI. Future Work

Using the insights gained from this study, there are several directions for further investigation to pursue.

Improving the ability to recognize facial expressions in the presence of occlusions can be achieved by fine-tuning existing models. The process of tweaking architectures or training strategies could lead to better handling situations where facial features are partially obscured. The aim would be to enhance accuracy, particularly for emotions that are more susceptible to misclassification.

Exploring the development of more advanced data augmentation techniques that are tailored for datasets with occlusions is another avenue worth exploring. By creating synthetic occluded images, we could enrich the training data and potentially improve the robustness of the models.

By exploring these paths, we could improve facial expression recognition and develop systems that are more accurate and reliable.

Acknowledgements

The authors are grateful to the Foundation for Science and Technology (FCT, Portugal) for financial support through national funds FCT/MCTES (PIDDAC) to CeDRI, UIDB/05757/2020 (DOI: 10.54499/UIDB/05757/2020) and UIDP/05757/2020 (DOI: 10.54499/UIDP/05757/2020) and SusTEC, LA/P/0007/2020 (DOI: 10.54499/LA/P/0007/2020).

References

[1] D. Matsumoto, M. Frank, and H. Hwang, *Nonverbal Communication: Science and Applications: Science and Applications* (EBSCO ebook academic collection). SAGE Publications, 2013, ISBN: 978-1-4129-9930-4. [Online]. Available: https://books.google.pt/books?id=PeOeu3qFFTIC.

[2] A. A. Pise, M. A. Alqahtani, P. Verma, *et al.*, "Methods for Facial Expression Recognition with Applications in Challenging Situations," *Computational Intelligence and Neuroscience*, vol. 2022, p. 9 261 438, May 2022. DOI: 10.1155/2022/9261438.

[3] N. Sarode and S. Bhatia, "Facial expression recognition," *International Journal on Computer Science and Engineering*, vol. 2, no. 5, pp. 1552–1557, 2010, Publisher: Citeseer.

[4] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion.," *Journal of Personality and Social Psychology*, vol. 17, no. 2, p. 124, 1971, Publisher: American Psychological Association.

[5] G. N. Foley and J. P. Gentile, "Nonverbal communication in psychotherapy," *Psychiatry (Edgmont)*, vol. 7, no. 6, pp. 38–44, Jun. 2010, Publisher: Libertas Academica.

[6] J. C. Lopes and R. P. Lopes, "A Review of Dynamic Difficulty Adjustment Methods for Serious Games," in *Optimization, Learning Algorithms and Applications*, A. I. Pereira, A. Koir, F. P. Fernandes, M. F. Pacheco, J. P. Teixeira, and R. P. Lopes, Eds., Cham: Springer International Publishing, 2022, pp. 144–159, ISBN: 978-3-031-23236-7.

[7] S. Alves, A. Marques, C. Queirós, and V. Orvalho, "LIFEisGAME prototype: A serious game about emotions for children with autism spectrum disorders.," *PsychNology Journal*, vol. 11, no. 3, 2013.

[8] B. Abirached, Y. Zhang, and J. H. Park, "Understanding user needs for serious games for teaching children with autism spectrum disorders emotions," in *EdMedia+ Innovate Learning*, Association for the Advancement of Computing in Education (AACE), 2012, pp. 1054–1063.

[9] M. Mascaró-Oliver, R. Mas-Sansó, E. Amengual-Alcover, and M. F. Roig-Maimó, "UIBVFED-Mask: A Dataset for Comparing Facial Expressions with and without Face Masks," en, *Data*, vol. 8, no. 1, p. 17, Jan. 2023, Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2306-5729. DOI: 10.3390/data8010017. [Online]. Available: https://www.mdpi.com/2306-5729/8/1/17 (visited on 08/17/2023).

[10] A. Rodrigues, J. Lopes, R. Lopes, and L. Teixeira, "Classification of Facial Expressions Under Partial Occlusion for VR Games," English, *Communications in Computer and Information Science*, vol. 1754 CCIS, pp. 804–819, 2022, ISBN: 9783031232350, ISSN: 1865-0929. DOI: 10.1007/978-3-031-23236-7_55.

[11] B. Yang, J. Wu, K. Ikeda, *et al.*, "Face-mask-aware Facial Expression Recognition based on Face Parsing and Vision Transformer," *Pattern Recognition Letters*, vol. 164, pp. 173–182, Dec. 2022, ISSN: 0167-8655. DOI: 10.1016/j.patrec.2022.11.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865522003312 (visited on 08/17/2023).

[12] L. Luo, D. Xue, and X. Feng, "EHANet: An Effective Hierarchical Aggregation Network for Face Parsing," *Applied Sciences*, vol. 10, no. 9, 2020, ISSN: 2076-3417. DOI: 10.3390/app10093135. [Online]. Available: https://www.mdpi.com/2076-3417/10/9/3135.

[13] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5549–5558.

[14] Y. Lin, J. Shen, Y. Wang, and M. Pantic, "RoI Tanh-polar transformer network for face parsing in the wild," *Image and Vision Computing*, vol. 112, p. 104 190, 2021, ISSN: 0262-8856. DOI: https://doi.org/10.1016/j.imavis.2021.104190. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885621000950.

[15] 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, no. 10, pp. 1499–1503, 2016. DOI: 10.1109/LSP.2016.2603342.

[16] B. Yang, J. Wu, and G. Hattori, "Facial Expression Recognition with the advent of face masks," in *Proceedings of the 19th International Conference on Mobile and Ubiquitous Multimedia*, ser. MUM '20, event-place: Essen, Germany, New York, NY, USA: Association for Computing Machinery, 2020, pp. 335–337, ISBN: 978-1-4503-8870-2. DOI: 10.1145/3428361.3432075. [Online]. Available: https://doi.org/10.1145/3428361.3432075.

[17] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society*

Conference on Computer Vision and Pattern Recognition - Workshops, 2010, pp. 94–101. DOI: 10.1109/CVPRW.2010.5543262.

[18] I. J. Goodfellow, D. Erhan, P. L. Carrier, *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, Springer, 2013, pp. 117–124.

[19] R. Magherini, E. Mussi, M. Servi, and Y. Volpe, "Emotion recognition in the times of COVID19: Coping with face masks," *Intelligent Systems with Applications*, vol. 15, p. 200 094, 2022, ISSN: 2667-3053. DOI: https://doi.org/10.1016/j.iswa.2022.200094. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667305322000345.

[20] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[21] A. Anwar and A. Raychowdhury, *Masked Face Recognition for Secure Authentication*, _eprint: 2008.11104, 2020.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[24] G. Castellano, B. De Carolis, and N. Macchiarulo, "Automatic facial emotion recognition at the COVID-19 pandemic time," en, *Multimedia Tools and Applications*, vol. 82, no. 9, pp. 12 751–12 769, Apr. 2023, ISSN: 1573-7721. DOI: 10.1007/s11042-022-14050-0. [Online]. Available: https://doi.org/10.1007/s11042-022-14050-0 (visited on 08/17/2023).

[25] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520. DOI: 10.1109/CVPR.2018.00474.

[26] O. Russakovsky, J. Deng, H. Su, *et al.*, *Imagenet large scale visual recognition challenge*, 2015. arXiv: 1409.0575 [cs.CV].

[27] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, *Bam: Bottleneck attention module*, 2018. arXiv: 1807.06514 [cs.CV].

[28] A. F. Abate, L. Cimmino, B.-C. Mocanu, F. Narducci, and F. Pop, "The limitations for expression recognition in computer vision introduced by facial masks," *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 11 305–11 319, 2023, ISSN: 1573-7721. DOI: 10.1007/s11042-022-13559-8. [Online]. Available: https://doi.org/10.1007/s11042-022-13559-8.

[29] L. Pham, T. H. Vu, and T. A. Tran, "Facial expression recognition using residual masking network," in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, 2021, pp. 4513–4519.

[30] N. Nischal, *Facial Expression Recognition with CNNs*, 2024. [Online]. Available: https://github.com/NJNischal/Facial-Expression-Recognition-with-CNNs.

[31] J. Shi, S. Zhu, and Z. Liang, *Learning to Amend Facial Expression Representation via De-albino and Affinity*, _eprint: 2103.10189, 2021.

[32] S. Li, W. Deng, and J. Du, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2584–2593. DOI: 10.1109/CVPR.2017.277.

[33] Prodesire, *Face-mask: Real-time face mask detection with OpenCV*, 2024. [Online]. Available: https://github.com/Prodesire/face-mask.

[34] L. Ruan, Y. Han, J. Sun, Q. Chen, and J. Li, "Facial expression recognition in facial occlusion scenarios: A path selection multi-network," *Displays*, vol. 74, p. 102 245, 2022, ISSN: 0141-9382. DOI: https://doi.org/10.1016/j.displa.2022.102245. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0141938222000701.

[35] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, IEEE, 1998, pp. 200–205.

[36] D. Lundqvist, A. Flykt, and A. Öhman, "Karolinska directed emotional faces," *PsycTESTS Dataset*, vol. 91, p. 630, 1998.

[37] A. Anwar and A. Raychowdhury, *Masked face recognition for secure authentication*, 2020. arXiv: 2008.11104 [cs.CV].

[38] A. V. Ikechukwu, S. Murali, R. Deepu, and R. Shivamurthy, "ResNet-50 vs VGG-19 vs training from scratch: A comparative analysis of the segmentation and classification of Pneumonia from chest X-ray images," *Global Transitions Proceedings*, vol. 2, no. 2, pp. 375–381, 2021, Publisher: Elsevier.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[40] Y. Jie, X. Ji, A. Yue, *et al.*, "Combined Multi-Layer Feature Fusion and Edge Detection Method for Distributed Photovoltaic Power Station Identification," *Energies*, vol. 13, p. 6742, Dec. 2020. DOI: 10.3390/en13246742.