Facial Empathy Analysis Through Deep Learning and Computer Vision Techniques in Mixed Reality Environments

Insaf Setitra[1], Domitile Lourdeaux[1] and Louenas Bounia[1,2]
[1]*UMR CNRS 7253 Heudiasyc, Sorbonne Université, Université de Technologie de Compiègne, 57 avenue de Landshut, Compiègne, France*
[2]*Université Sorbonne Paris Nord, Laboratoire d'Informatique de Paris-Nord (LIPN) - UMR-CNRS 7030, France*
{*isetitra, dlourdeaux, lbounia*}*@hds.utc.fr*

Keywords: Empathy, Facial Expression, Emotion Detection, Valence and Arousal.

Abstract: This paper introduces a novel approach for facial empathy analysis using deep learning and computer vision techniques within mixed reality environments. The primary objective is to detect and quantify empathic responses based on facial expressions, establishing the link between empathy and facial expressions. We propose the Deep Convolutional Neural Network with the Exponential Linear Unit activation function (ELU-DCNN). We moreover design an augmented reality platform with two main features (i). virtual overlay of a VR headset on the user's face and (ii). facial emotion recognition for users wearing the VR headset. Our target is to analyse facial expressions in immersed environments in order to assess the empathy of users while being immersed in specific environments. Our results analyse the feasibility and effectiveness of these models in detecting and quantifying empathy through facial expressions. This work contributes to the growing field of affective computing and highlights the potential of integrating advanced computer vision techniques in mixed reality applications to better understand human emotional responses.

1 INTRODUCTION

Empathy is a fundamental element of human interactions and plays a crucial role in communication and interpersonal relationships. Understanding and analyzing facial expressions associated with empathy can offer profound perspectives not only in the field of psychology, but also in various technological applications such as virtual reality, games and human-machine interactions.

The state of the art hence analyzes both empathy and facial expressions, with limited studies exploring their interrelation in the context of historical empathy. Some studies explore how empathy can improve human-centered design by understanding user needs and developing solutions accordingly (Zhu and Luo, 2023), (Somarathna et al., 2023), (Gareth W. Young and Smolic, 2022), (Ventura and Martingano, 2023), (Mathur et al., 2021), (Bang and Yildirim, 2018), (Shin, 2018). In contrast, our work diverges by concentrating on the use of computer vision and deep learning techniques to analyze empathic responses in historical scenarios, aiming to adapt and enrich these scenarios through advanced technological approaches. The main goal is to create a system that can detect and analyze facial expressions to measure empathy. The project contributes to the improvement of virtual reality scenarios in a historical setting (mainly in the memorial of Compiègne) based on detected empathy.

1.1 Empathy Detection

Empathy is divided into three distinct components: affective, cognitive and associative empathy (Shen, 2010), (Ventura and Martingano, 2023). Affective empathy is the ability to feel the emotions of another person, cognitive empathy is the ability to understand the thoughts and feelings of others, and associative empathy combines the emotional and cognitive aspects, allowing individuals to put themselves completely in the other's place. In our study, we are mainly interested in cognitive empathy. Indeed, the purpose of visiting a historical museum is to understand historical events without directly feeling what the individuals (deportees for instance) could have felt. The study in (Hasan et al., 2024) revealed that while empathy detection systems use various types of signals, there is a predominance of empathy analysis from texts as opposite to facial expressions. Question-

Setitra, I., Lourdeaux, D. and Bounia, L.
Facial Empathy Analysis Through Deep Learning and Computer Vision Techniques in Mixed Reality Environments.
DOI: 10.5220/0013057800003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 31-39
ISBN: 978-989-758-737-5; ISSN: 2184-433X

naires allow to label the collected data. The Toronto empathy questionnaire (Spreng et al., 2009) is frequently cited in the literature. Shin (Shin, 2018) proposed a model examining the relationships between immersion, presence, flow, embodiment and empathy in a virtual environment. Bang and Yildirim (Bang and Yildirim, 2018) conducted a study to assess the effectiveness of virtual reality storytelling in building user empathy. To do this, they compared two groups of participants: the first watched the documentary After Solitary[1] using a VR headset, while the second one watched the same video in 360^o format on a desktop computer. In the context of measuring empathy, researchers used the State Empathy Questionnaire (SEQ)(Shen, 2010). Leena et al.(Mathur et al., 2021) conducted an experiment to collect a new set of empathy data in an innovative interaction context, where participants listened to stories told by the storyteller robot LuxAI. After listening to the three stories, participants completed the SEQ Questionnaire to assess their empathic reactions. Video information of participants was first extracted by OpenFace 2.0 (Baltrusaitis et al., 2018) to identify certain facial features, including expression movement, eye angle and head position. (Gareth W. Young and Smolic, 2022) states that a good immersion creates an "illusion of body exchange" that facilitates the adoption of the perspective of the incarnate character. Among scenarios that evoke empathy stands The Last Goodbye[2], a virtual reality experience that allows users to visit the remains of a Nazi concentration camp in the company of a Holocaust survivor, Pinchas Gutter. This scenario is intended to evoke intense emotions such as sadness, anger and deep reflection. "VR World War II" [3] is another virtual reality scenario that immerses users in the events of World War II. In (Xue et al., 2023)[4], a database is presented which includes 73 extracts that induce valence variations and emotional activation. A similar dataset is presented in (Li et al., 2017)[5] with a set of 360^o videos. AVDOS-VR (Gnacek et al., 2024)[6] contains 30-second videos with activation and valence information evaluated at each second. Arousal and valence (positive/negative) are often used to evaluate induced emotions. Self-assessment by participants via the SAM scale is a widespread method. Correlations are observed between head upward movements and high activation levels (Somarathna et al., 2023). A subset of AVDOS-VR is also presented in (Xue et al., 2023).

In our study we choose four scenarios from AVDOS-VR (Gnacek et al., 2024) that are used using a desktop application we developed and two VR scenarios (360^o) from (Li et al., 2017). We use the Toronto questionnaire (Spreng et al., 2009) before the experiment and the SEQ questionnaire (Shen, 2010) after the experience. We provide more details in the following sections.

1.2 Facial Expression Recoginition

According to Rakibul et al. (Hasan et al., 2024), for empathy detection in the deep learning category, models based on Convolutional Neural networks (CNN) and Recurrent Neural Networks (RNN) are most frequently used, while in the classical machine learning category, SVM is the most frequently used. Traditionally, seven basic emotions are classified: fear, anger, disgust, happiness, neutrality, sadness and surprise while the main datasets used in the literature are FER2013, AffectNet, CK+, eNTERFACE'05 (Martin et al., 2006). Li and Deng (Li and Deng, 2022) present a comprehensive review on Facial Expression Recognition (FER) and describes the standard process of a deep FER system, including preprocessing, learning of deep characteristics and classification. (Mohamed et al., 2022) presents a method based on the association of a pre-trained CNN models (VGG16, ResNet50) used as a feature extractor with a Multi-Layer Perceptron (MLP) classifier. (Demochkina and Savchenko, 2021) presents a method to recognize facial expressions in videos, using the proposed MobileEmotiFace netword. Kas et al. (Kas et al., 2021) presents a framework that combines texture and shape characteristics from 49 landmarks detected on a facial image. The shape is extracted using Histogram of Oriented Gradient (HOG) and the texture using an Orthogonal and Parallel-based Directions Generic Quad Map Binary Patterns (OPD-GQMBP). Ezati et al. in (Ezati et al., 2024) highlight the challenges posed by high computational complexity and variations of multi-view poses in real-

[1] https://www.youtube.com/watch?v=G7_YvGDh9Uc&t=14s

[2] USC Shoah Foundation (2020, Septembre 18). The last GOODBYE [VR documentary]. Gabo Arora and Ari Palitz. https://sfi.usc.edu/lastgoodbye

[3] World War II Foundation (2021-2024) VR video series. https://www.youtube.com/playlist?list=PL2A7-aRM5qjU7KKRIdL-LsPObYfZHT7Od

[4] https://www.dis.cwi.nl/ceap-360vr-dataset/, https://github.com/cwi-dis/CEAP-360VR-Dataset/tree/master?tab=readme-ov-file

[5] https://vhil.stanford.edu/public-database-360-videos-corresponding-ratings-arousal-and-valence

[6] https://github.com/michalgnacek/AVDOS-VR/tree/main , https://www.gnacek.com/affective-video-database-online-study

world contexts and propose a Lightweight Attentional Network incorporating Multi-Scale Feature Fusion (LANMSFF). Ma et al (Ma et al., 2024) propose the FER-YOLO-Mamba model, which integrates the principles of Mamba and YOLO technologies to facilitate efficient recognition and localization of facial expressions. Sun et al. (Sun et al., 2023) propose a new self-supervised approach called SVFAP that uses self-supervised learning to overcome challenges related to overfitting and the high cost of creating datasets. Another significant area of research examines how different facial regions contribute to the recognition of facial expression. The observations of Wegrzyn et al. (Wegrzyn et al., 2017) show that the lower part of the face is about joy and disgust while the upper part informs about anger, Fear, surprise and sadness. Wingenbach (Wingenbach, 2023) shows the relationship between facial muscles and facial expression. Using facial electromyography[7], it is possible to detect the slightest muscle contractions and identify the characteristic Action Units (AUs). In addition, it is possible to combine these AUs to determine an emotion. The AU method has been used in several other studies such as (Yao et al., 2021). Huc et. al. (Huc et al., 2023) studies the effects of emotional attribution errors on people with or without masks. This work shows that the bottom of the face contributes to the identification of emotions of joy, sadness and surprise while the top of the face allows to recognize fear which contradicts somewhat the (Wingenbach, 2023). Other confusions such as the confusion of fear and surprise, anger and disgust and sadness and fear and all with the neutral emotion are also described. (Poux et al., 2020) is based on the propagation of facial movement to overcome difficulties related to occlusions. (Minaee et al., 2021) presents an innovative approach for facial expression recognition based on a convolutional attentional network with a spatial attention mechanism. To validate the attentional approach, a saliency map of important regions is generated. The results confirmed that different expressions are indeed sensitive to different parts of the face, for example the mouth for joy and the eyes for anger.

2 OUR APPROACH

Our approach to facial empathy analysis is structured into three key parts. The first part (Subsection 2.1) focuses on facial expression recognition. We employ a proposed neural network that fully predicts the expression from an image. This network is combined with various feature extractors, including pixels, Histogram of Oriented Gradients (HOG), MobileNet, and VGG, along with classification algorithms such as K-Nearest Neighbors (KNN) and Support Vector Machines (SVM). In the second part (Subsection 2.2), we introduce a novel aspect of our dataset by adding a mask to the expressions. This modification simulates an environment where individuals are immersed in Virtual Reality (VR), allowing us to analyze their facial expressions under these conditions. The last part (Subsection 3.2) outlines our experimental setup. We describe a series of scenarios designed to evoke empathy, during which we capture video recordings of participants' reactions, both with and without wearing a VR headset. Participants complete during the experimentation empathy questionnaires to provide additional insights into their empathic responses. This comprehensive approach enables us to analyze and understand the relationship between facial expressions and empathy in both traditional and VR environments.

2.1 Facial Expression Recognition

The network we propose to classify facial expressions is inspired of the work of Debnath et al.(Debnath et al., 2022). Our proposed model consists of six convolution layers organized into three blocks:

- The first convolution block contains two layers using 5 filters with 64 filters each.
- The second block consists of two layers using 3×3 filters with 128 filters.
- The third block has two convolution layers using 3×3 filters with 256 filters.
- Each convolution layer is followed by batch normalization which helps to stabilize and accelerate the drive. Moreover, an ELU (Exponential Linear Unit) activation function is also applied after each layer to improve network convergence.
- The model integrates MaxPooling layers after each convolution layer block to reduce dimensionality.
- Dropout layers prevent overfitting by ensuring regularization as the number of parameters increases.
- Flatten layer is finally used to convert the feature maps into a one-dimensional vector followed by a dense layer of 128 neurons with ELU activation and batch normalization.
- The output layer uses a softmax activation for multi-class classification.

[7] A medical technique that studies the function of nerves and muscles

In addition, to improve the diversity and balance of our training dataset, we have used a data augmentation strategy shown in Figure 1. The set of augmentations used are as follows given that x, y and x', y' are the image coordinates before and after the augmentation respectively:

- **Gaussian blur**: The weighted average of the neighboring pixels for each pixel in the image is applied. We use the Gaussian kernel to do so ($G(x,y) = \frac{1}{2\pi\sigma^2} exp^{\frac{-x^2+y^2}{2\sigma^2}}$).
- **Affine transformation**: The combination of translations, rotations, scaling and shears (bias distortions) is applied $\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}$.
- **Euclidean Transformation**: includes only rotations and translations $\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix}$ with Where θ is the angle of rotation and (t_x, t_y) is the translation vector.
- **Total Transformation**: transformation of the quadrilaterals into other quadrilaterals $\begin{pmatrix} x' \\ y' \\ w' \end{pmatrix} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$.
- **Contrast Modification:** Higher contrast makes light areas brighter and dark areas darker $I\prime(x,y) = \alpha I(x,y) + \beta$ with I the image intensity and α, β the contrast and brithness factors respectively.
- **Image Flipping (flip):** either horizontal ($x = -x, y = y$) or vertical ($x = x, y = -y$).

For the feature based facial expression recognition, we extract 4 types of features and use two main classifiers. The feature extractors are described briefly in the following:

- **Pixels:** each image is flattened into a vector of pixels. Then, this vector is normalized to have an average of 0 and a standard deviation of 1, which improves the performance of machine learning algorithms.
- **Histogram of Oriented Gradient HOG:** he histograms of all cells are grouped together and used as a feature vector.
- **VGG Feature Extractor:** in VGG16 (Simonyan and Zisserman, 2014) we adopt the last layer (4096 dimensional vector) as the feature vector without any training.
- **Mobile Net Feature Extractor:** Similarily to VGG, we use the last layer of MobileNet (Howard et al., 2017) as a feature vector.

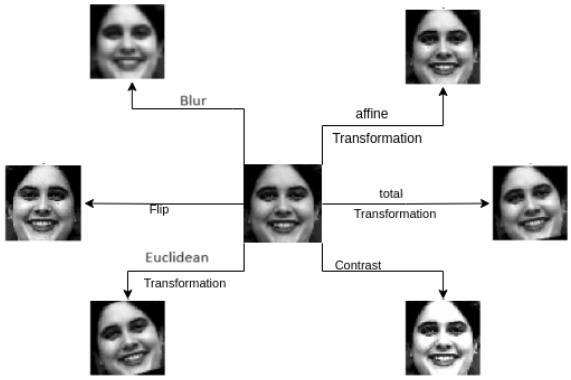

Figure 1: Examples of the used augmentations.

Finally we use SVM and KNN for classification and use in our approach the cross-validation to determine the optimal number of neighbors.

2.2 Augmented Reality for Facial Expression Recognition in Immersion

Virtual reality headset overlay on faces via Augmented Reality (AR) offers an innovative approach to the study of emotion recognition in the context of experiences in emotion recognition. This work also allows to assess the potential impact of facial occlusion by physical devices on the recognition of emotions.

The AR algorithm relies on two pre-prepared VR headset images: a front view and a profile (sagittal) view. The first crucial step in the process is to analyze the input image to detect and understand the face geometry. This phase is fundamental because it provides the necessary information to position and orient the virtual VR headset correctly. The algorithm operates in the following stages. (i) **Face Detection**:This process a machine learning-based method known as BlazeFace (Bazarevsky et al., 2019), a lightweight and efficient face detector designed for real-time applications. BlazeFace leverages a single-shot detection (SSD) architecture that quickly identifies the bounding box of the face within the image, enabling real-time face detection with high precision. (ii). **Face Mesh Generation**: Within the Region Of Interest (ROI), a mesh grid is generated, mapping out the facial structure. This grid consists of numerous points, known as landmarks, strategically positioned to capture critical facial features such as the eyes,

nose, and mouth. (iii). **Landmark Localization**: The algorithm uses the convolutional neural network MobileNetV2 (Sandler et al., 2019) to predict the precise locations of these landmarks. (iv). **Temporal Filtering**: For real-time applications, temporal filtering is applied to stabilize the landmark positions across consecutive frames. This reduces jitter and ensures smooth tracking of facial movements.

From the set of detected landmarks, we selects four specific points that will serve as a reference for positioning the VR headset. These points are strategically chosen to frame the area where the helmet will be placed, usually around the eyes and temple and determine the geometric transformation applied to the helmet. To determine the pitch, roll, and yaw angles of the head, the pose can be calculated from 3D-2D point correspondences using relevant algorithms and facial landmarks. This problem involves solving for the rotation (r) and translation (t) that minimize the projection error from 3D-2D point correspondences. The rotation vector r represents the axis of rotation in 3D space, and its magnitude represents the angle of rotation. To convert this rotation vector into a rotation matrix R, we use the Rodrigues' rotation formula (Hartley and Zisserman, 2003). The Rodrigues' rotation formula converts a rotation vector into a rotation matrix through the following steps:

- Compute the angle of rotation θ as the magnitude of the rotation vector: $\theta = \|r\|$.
- Compute the unit vector $k = \frac{r}{\theta}$
- Construct the skew-symmetric cross-product matrix K of k : $K = \begin{bmatrix} 0 & -k_z & k_y \\ k_z & 0 & -k_x \\ -k_y & k_x & 0 \end{bmatrix}$
- Compute the rotation matrix R as $R = I + sin(\theta)K + (1 - cos(\theta))K^2$ with I the identity matrix.
- Finally, compute the Euler angles from the rotation matrix R. The Euler angles (roll, pitch, and yaw) describe the orientation of an object in three-dimensional space. These angles are extracted from the rotation matrix using as follows:
  - Roll α : $\alpha = atan(R_{32}, R_{33})$
  - Pitch β : $\beta = atan(-R_{31}, \sqrt{R_{32}^2 + R_{33}^2})$
  - Yaw γ : $\gamma = atan(R_{21}, R_{11})$
  - $R = \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{bmatrix}$ being the rotation matrix.

Once calculated, the transformation matrix is applied to the helmet image. This distorts the original image according to the set parameters, producing a version of the headset aligned with the face in the image.

2.3 Immersion Protocol and Empathy Analysis

For the study of reactions, we have chosen four short 2D videos from AVDOS-VR (Gnacek et al., 2024) and two virtual reality videos from (Li et al., 2017). The scenarios are as follows:

- **Police Helicopter Captures Armed Confrontation** (vidéo 13 of AVDOS)**:** 2D video of 30 seconds showing a police drill where an officer comes to the aid of a colleague who was shot 2.659 and 6.652 for valence and arousal respectively.
- **Sick Boy Crying During an Interview** (vidéo 19 of AVDOS)**:** 2D video of 30 seconds showing an excerpt from an interview with a child suffering from a serious illness 2.434 and 5.623 for valence and arousal respectively.
- **Soldiers Marching and Singing a Pop Song** (vidéo 56 of AVDOS)**:** 2D video of 30 seconds showing a group of soldiers parading by singing the song "Barbie Girl" 6.283 and 5.975 for valence and arousal respectively.
- **Toddler Laughing at Torn Paper Pages** (vidéo 51 of AVDOS)**:** 2D video of 30 seconds showing a young child laughs when an adult tears up a piece of paper with 7.07 and 6.429 for valence and arousal respectively.
- **Survive a Bear Attack in VR:** a 360^o video of 90 seconds showing a bear approaching dangerously to a group of 3 campers, who decide to run away in their car after having distracted the bear with a cookie with 5.22 and 5 for valence and arousal respectively.
- **Solitary Confinement:** a 360^o video of 221 seconds that puts the viewer in the shoes of an inmate in isolation, while listening to a testimony from a former prisoner 2.38 and 4.25 for valence and arousal respectively.

Moreover, in the experimental protocol, we have developed a Graphical User Interface (GUI) to allow visualization the videos while simultaneously capturing the webcam video stream. We use the Toronto questionnaire (Spreng et al., 2009) before the experiment to assess the empathy of the participant, followed by the SEQ questionnaire (Shen, 2010) after the experience. In order to collect responses of the questionnaires, we use Google Forms in which we used the same questions as the questionnaires. We

also add for the questionnaire a precision about which individual in the video to rely for the empathic response. For example, in the "Police helicopter captures armed confrontation" scenario, we ask the participant to choose which actor in the video to relate to (the first policeman, the second, or the confronting actor). More specifically, we provide the participants with the following information:

- **For the Preliminary Phase, Follow the Following Steps:**
  - Inform the participant that they will be viewing a series of videos, each followed by a questionnaire. Also mention that some videos may contain graphic scenes (notably video 13).
  - Provide participants with contentment forms to be filled, signed and returned.
  - Place the participant in front of a computer in a quiet room with a neutral background.
  - Ask the participant to complete the Toronto Questionnaire.
  - Explain the operation of the graphical interface used to view videos.
- **For each 2D Video, Repeat the Following Steps:**
  - The experimental staff leaves the room.
  - The participant starts the video using the GUI.
  - At the end of the video, the experimental staff returns to the room and asks the participant to complete the State Empathy Questionnaire corresponding to the video.
- **For each 3D** (360^o) **Video, Repeat the Following Steps:**
  - The experimentation staff prepares the headset with the video ready to be launched.
  - The experimental staff leaves the room.
  - At the end of the video, the experimental staff returns to the room and asks the participant to complete the State Empathy Questionnaire corresponding to the video.

After the experiment, we annotate the captured videos with the appropriate facial expression. We mainly select a representative image of the expressions and reactions of the person being filmed, and annotate the emotion observed among the following seven labels: anger, disgust, fear, happiness, sadness, surprise and neutrality. Finally, we crop the image to retain only the participant's face.

3 EXPERIMENTS

3.1 Facial Expression Recognition Experiments

Experiments for facial expression recognition are conducted on the Facial Expression Recognition 2013 (FER2013) database (Goodfellow et al., 2013). The database includes 35.887 grayscale images of faces with dimensions of 48×48 pixels. The images are labeled in seven categories of emotions: anger, disgust, fear, happiness, sadness, surprise and neutral. Although FER2013 is particularly useful for training and validating facial recognition models due to its large size and the diversity of emotions represented, classes in the database are not balanced. Particularly the class 'disgust' has more than 16 times fewer samples than the class 'happiness'. Figure 2 highlights the imbalance in the distribution of emotion classes. We therefore applied data augmentation described previously in order to balance the classes. 10.000 images were obtained for each class. We summarize in Table 1 the different settings along with the obtained accuracy. For our ELU-DCNN, the Dropout layer has a rate of 0.6 to avoid overfitting and the Adam optimizer is used. We used early stop callbacks (EarlyStopping) to monitor the accuracy validation metric, with a criterion of patience of 11 epochs and a restoration of the best weights. Finally, we used a scheduler of reduction of the learning rate (ReduceLROnPlateau) in order to reduce the learning rate by half after 7 epochs without improvement. The model was trained for 100 epochs with lots (batch) of 32.

Table 1: Results of facial expression recognition using our approach on the FER2013 dataset.

Approach	Details	Accuracy
Basic approach	batch size = 64 Early stopping with p=5 Imbalanced classes	0.63
Data augmentation	10.000 images par classes	0.69
Reduced batch size	batch size = 32 Early stopping with p=11 Nadam optimizer	0.78
Parameter refinement	Adam Optimizer	0.81

As can be seen from the table, we obtain an accuracy of 81% on our test data sample. Figure 3 also shows that the accuracy is increasing and that it eventually stabilizes around 80% after the 40^{th} epoch. Finally Figure 4 shows the confusion matrix for the validation set. As can be seen, the model predicts in a similar manner the different classes of expressions. Hence, our model predicts 74% anger, 97% disgust, 68% fear, 85% happiness, 72% sadness, 90% surprise and 77% neutral.

Emotion	# images
0 'happiness'	8989
1 'neutral'	6198
2 'sadness'	6077
3 'fear'	5121
4 'anger'	4953
5 'surprise'	4002
6 'disgust'	547

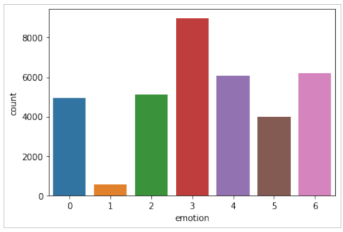

Figure 2: Distribution of Emotion Classes in the Dataset: (left) Table of Emotion Names and Image Counts with (right) Corresponding Histogram Illustrating Class Imbalance.

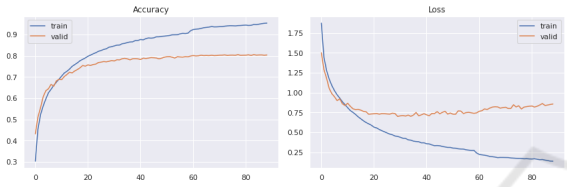

Figure 3: Accuracy (left) and loss (right) with respect to number of epochs (x-axis).

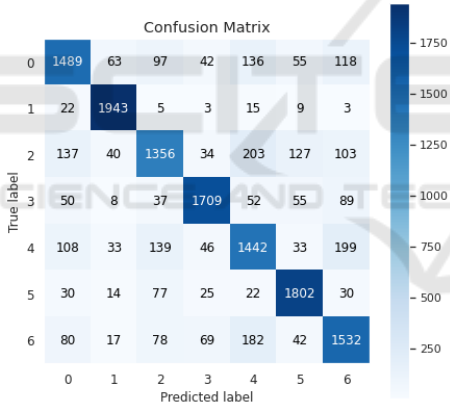

Figure 4: Confusion matrix for the validation set of FER2013 dataset using our approach.

3.2 Empathy Analysis Experiments

In order to retrain our model on occluded faces and hence simulate an environment where individuals are immersed, we apply the augmented reality transformations on the FER dataset. This has the advantage of training a facial expression classifier without the need to re-annotate the data (as FER2013 is already annotated with facial expression classes). However, the accuracy drops to 67%. Some emotions without the information from the top of the head struggle to be recognized because of the lack of information. Some emotions, such as anger, are strongly expressed in the eye and eyebrow area, areas typically masked by a VR headset. This occlusion can significantly reduce the accuracy of emotional recognition.

We simultaneously performed the test protocol for empathy analysis as described in Subsection . To summarize, the protocol consists on:

- Ask participants to sign the consent form to use their videos for this research.
- Ask participants to fill the Toronto form.
- Show scenarios (either 2D or 360^o videos) to the participants and capture their videos while they watch the videos.
- After watching each scenario, ask participants to fill the empathy SEQ quetionnaire that we reproduced in the google form.
-

By the end of the tests, We obtained a set of faces of 8 individuals who participated in the experiments. Each face was associated with emotions that we labelled manually and on which we were able to test our different models of facial expression recognition. Our classifier has achieved an accuracy of 20%. The model correctly identified the expression of happiness but failed to recognize the expression of surprise, wrongly classifying it as anger. The main limitations are glasses on the face, poor image quality and different angles. Since the model did not perform to a high extent on the captured videos, we analyzed empathy based on the facial expressions that we manually annotated. We used the results of the questionnaires to assess the level of empathy associated with different emotions. As illustrated in Figure 5, different emotions correspond to varying levels of empathic resonance.

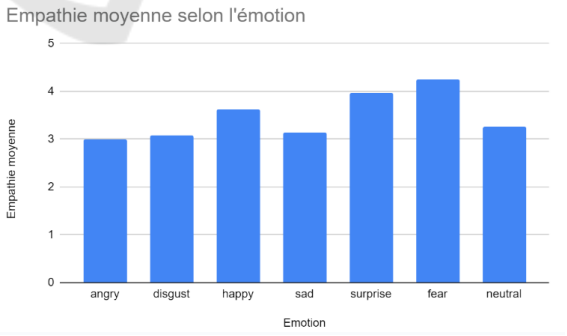

Figure 5: Confusion matrix for the validation set of FER2013 dataset using our approch.

It is clear that fear and surprise have the highest empathy scores, while anger has the lowest score. However, the differences in scores are not particularly marked. This situation can be explained by several factors. We depict some of them in the following:

- **Participants' Familiarity with the Videos:** Many participants had already participated in the selection of the videos, their familiarity with video content may lessen the intensity of their empathic response, as they know what to expect and are less surprised by the emotions depicted.
- **Data Set Limitation:** Our data set is still limited in terms of diversity and quantity.
- **Individual Variability:** Empathic responses can vary greatly from person to person based on individual factors such as personal experiences, emotional sensitivity, and innate empathic skills.
- **Nature of Emotions:** Some emotions can naturally trigger stronger empathic reactions. For example, fear and surprise are intense and often immediate emotions, which may explain why they score high. Other emotions, such as anger, can be more complex and cause various reactions, some people may feel empathy while others feel resistance or rejection.

In order to overcome these limitations and achieve more representative and generalizable results, we can propose the following improvements. (i). **Increase the scale of the experience**: By expanding the number of participants and diversifying demographic groups, we can get a better representation of the empathic reactions. (ii). **Diversification of emotional stimuli**: By using a greater variety of videos and images to represent emotions, we can reduce the familiarity effect and capture a wider range of reactions. (iii). **Improvement of Evaluation Conditions:** By standardizing the viewing conditions and minimizing distractions, we can ensure that the reactions of participants are as natural and authentic as possible. (iv). **Use of Advanced Empathy Measurement Techniques:** By integrating psychophysiological measures such as eye movement tracking, By analyzing facial expressions in real time, and monitoring physiological responses, we can obtain more accurate and objective data on levels of empathy.

By implementing these strategies, we hope to gain more accurate and reliable insights into the levels of empathy associated with different emotions, and thus improve our understanding of the mechanisms of empathetic resonance.

4 CONCLUSION

In this work, we developed a comprehensive system for facial empathy analysis using computer vision and machine learning techniques. Our experimentation setup included scenarios that evoke empathy, capturing participants' facial reactions through video recordings, and subsequently measuring their empathy levels using questionnaires. The findings from our experiments indicate at some extent the correlation between specific facial expressions and empathic responses. Overall, this work contributes to the field by bridging the gap between facial expression analysis and empathy detection, offering a novel approach that can be applied in various domains, including psychology, human-computer interaction, and virtual reality. Future research could focus on refining the models and algorithms for even greater accuracy and exploring additional applications of this technology in real-world scenarios.

REFERENCES

Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L.-P. (2018). Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*, pages 59–66.

Bang, E. and Yildirim, C. (2018). Virtually empathetic?: Examining the effects of virtual reality storytelling on empathy. In Chen, J. Y. and Fragomeni, G., editors, *Virtual, Augmented and Mixed Reality: Interaction, Navigation, Visualization, Embodiment, and Simulation*, pages 290–298, Cham. Springer International Publishing.

Bazarevsky, V., Kartynnik, Y., Vakunov, A., Raveendran, K., and Grundmann, M. (2019). Blazeface: Submillisecond neural face detection on mobile gpus.

Debnath, T., Reza, M., Rahman, A., Beheshti, A., Band, S., and Alinejad-Rokny, H. (2022). Four-layer convnet to facial emotion recognition with minimal epochs and the significance of data diversity. *Scientific Reports*, 12:1–18. Copyright the Crown 2022. Version archived for private and non-commercial use with the permission of the author/s and according to publisher conditions. For further rights please contact the publisher.

Demochkina, P. and Savchenko, A. V. (2021). Mobileemotiface: Efficient facial image representations in video-based emotion recognition on mobile devices. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part V*, page 266–274, Berlin, Heidelberg. Springer-Verlag.

Ezati, A., Dezyani, M., Rana, R., Rajabi, R., and Ayatollahi, A. (2024). A lightweight attention-based deep network via multi-scale feature fusion for multi-view facial expression recognition. *ArXiv*, abs/2403.14318.

Gareth W. Young, N. O. and Smolic, A. (2022). Exploring virtual reality for quality immersive empathy building experiences. *Behaviour and Information Technology*, 41(16):3415–3431.

Gnacek, M., Quintero, L., Mavridou, I., Balaguer-Ballester, E., Kostoulas, T., Nduka, C., and Seiss, E. (2024).

AVDOS-VR: Affective Video Database with Physiological Signals and Continuous Ratings Collected Remotely in VR. *Scientific Data*, 11(1).

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Chuang, Z., and Bengio, Y. (2013). Challenges in representation learning: A report on three machine learning contests. In Lee, M., Hirose, A., Hou, Z.-G., and Kil, R. M., editors, *Neural Information Processing*, pages 117–124, Berlin, Heidelberg. Springer Berlin Heidelberg.

Hartley, R. and Zisserman, A. (2003). *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2 edition.

Hasan, M. R., Hossain, M. Z., Ghosh, S., Krishna, A., and Gedeon, T. (2024). Empathy detection from text, audiovisual, audio or physiological signals: Task formulations and machine learning methods.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications.

Huc, M., Bush, K., Atias, G., Berrigan, L., Cox, S., and Jaworska, N. (2023). Recognition of masked and unmasked facial expressions in males and females and relations with mental wellness. *Frontiers in Psychology*, 14.

Kas, M., merabet, Y. E., Ruichek, Y., and Messoussi, R. (2021). New framework for person-independent facial expression recognition combining textural and shape analysis through new feature extraction approach. *Information Sciences*, 549:200–220.

Li, B. J., Bailenson, J. N., Pines, A., Greenleaf, W. J., and Williams, L. M. (2017). A public database of immersive vr videos with corresponding ratings of arousal, valence, and correlations between head movements and self report measures. *Frontiers in Psychology*, 8.

Li, S. and Deng, W. (2022). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3):1195–1215.

Ma, H., Lei, S., Celik, T., and Li, H.-C. (2024). Fer-yolo-mamba: Facial expression detection and classification based on selective state space.

Martin, O., Kotsia, I., Macq, B., and Pitas, I. (2006). The enterface' 05 audio-visual emotion database. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pages 8–8.

Mathur, L., Spitale, M., Xi, H., Li, J., and Matarić, M. J. (2021). Modeling user empathy elicited by a robot storyteller. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8.

Minaee, S., Minaei, M., and Abdolrashidi, A. (2021). Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, 21(9).

Mohamed, B., Daoud, M., Mohamed, B., and taleb ahmed, A. (2022). Improvement of emotion recognition from facial images using deep learning and early stopping cross validation. *Multimedia Tools and Applications*, 81.

Poux, D., Allaert, B., Mennesson, J., Ihaddadene, N., Bilasco, I. M., and Djeraba, C. (2020). Facial expressions analysis under occlusions based on specificities of facial motion propagation. *Multimedia Tools and Applications*, 80(15):22405–22427.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2019). Mobilenetv2: Inverted residuals and linear bottlenecks.

Shen, L. (2010). On a scale of state empathy during message processing. *Western Journal of Communication*, 74:504–524.

Shin, D. (2018). Empathy and embodied experience in virtual environment: To what extent can virtual reality stimulate empathy and embodied experience? *Computers in Human Behavior*, 78:64–73.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Somarathna, R., Bednarz, T., and Mohammadi, G. (2023). Virtual reality for emotion elicitation – a review. *IEEE Transactions on Affective Computing*, 14(4):2626–2645.

Spreng, R. N., Mckinnon, M., Mar, R., and Levine, B. (2009). The toronto empathy questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of personality assessment*, 91:62–71.

Sun, L., Lian, Z., Wang, K., He, Y., Xu, M., Sun, H., Liu, B., and Tao, J. (2023). Svfap: Self-supervised video facial affect perceiver.

Ventura, S. and Martingano, A. J. (2023). Roundtable: Raising empathy through virtual reality. In Ventura, S., editor, *Empathy*, chapter 3. IntechOpen, Rijeka.

Wegrzyn, M., Vogt, M., Kireclioglu, B., Schneider, J., and Kissler, J. (2017). Mapping the emotional face. how individual face parts contribute to successful emotion recognition. *PLOS ONE*, 12.

Wingenbach, T. S. H. (2023). *Facial EMG – Investigating the Interplay of Facial Muscles and Emotions*, pages 283–300. Springer International Publishing, Cham.

Xue, T., Ali, A. E., Zhang, T., Ding, G., and Cesar, P. (2023). Ceap-360vr: A continuous physiological and behavioral emotion annotation dataset for 360° vr videos. *IEEE Transactions on Multimedia*, 25:243–255.

Yao, L., Wan, Y., Ni, H., and Xu, B. (2021). Action unit classification for facial expression recognition using active learning and svm. *Multimedia Tools and Applications*, 80.

Zhu, Q. and Luo, J. (2023). Toward artificial empathy for human-centered design: A framework.